TECHNISCHE UNIVERSITÄT DRESDEN

Proceedings of the 3rd International Conference on Models and Technologies for Intelligent Transportation Systems 2013

2.–4. December 2013, Dresden

This proceedings are published in:
Proceedings of the 3rd International Conference on Models and Technologies for Intelligent Transportation Systems 2013, Ed. by T. Albrecht, B. Jaekel, M. Lehnert. Verkehrstelematik. Vol. 3. Dresden: TUDpress, 2013.

ISBN: 978-3-944331-34-8

Preface

Often, the term “intelligent transportation systems” (ITS) is associated with road traffic only. We at the Institute of Traffic Telematics of the “Friedrich List” Faculty for Transport and Traffic Sciences of Technische Universität Dresden have always seen it in a broader way: each transport mode should be regarded as an intelligent transportation system itself, but also as part of *one* intelligent transportation system with “intelligent” intramodal and intermodal interfaces. Only such an “intelligent” transportation system will be able to meet the challenges of increasing traffic demand, limited resource availability and growing quality expectations of the customers.

A similar concept is followed by the bi-annual series of conferences “Models and Technologies for Intelligent Transportation Systems” (2009 in Rome, 2011 in Leuven). It provides a forum for excellent scientific contributions with an intensive exchange between theory and practice and the presentation of case studies for all transport modes. We applied to host the third conference in the series to give a discussion forum for control engineers, computer scientists, mathematicians and other researchers and practitioners about the variety of traffic management problems that can be solved using similar methods and technologies, but with application specific models, objective functions and constraints.

This book comprises fifty short papers accepted for presentation at the Third International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS 2013), which has been held in Dresden in December 2013. All submissions have undergone intensive reviews by the organisers of the special sessions, the members of the scientific and technical advisory committees and further external experts in the field. We are grateful for their voluntary work.

As the conference itself, the proceedings are structured in twelve streams: the more model-oriented streams of Road-Bound Public Transport Management, Modelling and Control of Urban Traffic Flow, Railway Traffic Management in four different sessions, Air Traffic Management, Water Traffic and Traffic and Transit Assignment, as well as the technology-oriented streams of Floating Car Data, Localisation Technologies for Intelligent Transportation Systems and Image Processing in Transportation.

With this broad range of topics the book will be of interest to a number of groups: ITS experts in research and industry, students of transport and control engineering, operations research and computer science. The case studies will also be of interest for transport operators and members of traffic administration.

We wish to thank all the authors for devoting their time and energy to preparing their excellent papers and for submitting them to the MT-ITS 2013 conference.

Furthermore we wish to express our gratitude to the financial support of the conference by industry sponsors, the German Academic Exchange Service (DAAD) and the Deutsche Forschungsgesellschaft (DFG). Without it, the publication of the conference proceedings as a book would not have been possible.

Thomas Albrecht (Conference chair), Birgit Jaekel and Martin Lehnert
The Editors, Dresden, 2013

Table of Contents

Modelling and Control of Urban Traffic Flow
(Stefan Lämmer, TU Dresden)

A Microscopic Simulation Approach for Optimization of Taxi Services **1**
(Michał Maciejewski, Kai Nagel)

Challenges for Better Understanding and Simulating Urban Traffic - the Zurich Experience **11**
(Christian Heimgartner, Monica Menendez)

Inflow-Regulating Traffic Light Control to Avoid Queue-Spillovers in Urban Road Networks **23**
(Stefan Lämmer, Martin Treiber, Markus Rausch)

Integrating Weather Impact in Travel Demand Models for Private Motorised Transport **35**
(Bernhard Heilmann, Martin Reinthaler, Johannes Asamer, Lena Fehrenbach, Juliane Pillat, Jochen Lohmiller, Markus Friedrich, Karl Schedler)

Method for the Organization of Daily Activity Chains **47**
(Domokos Esztergár-Kiss, Dénes Válóczi)

Minimization of Vehicle Stops by an Early Termination of Green Times in Traffic-Light Controlled Road Networks **57**
(Kathleen Tischler, Stefan Lämmer)

Using Nanoscopic Simulations to validate the Benefit of Advanced Driver Assistance Systems in complex Traffic Scenarios **69**
(Torsten Schubert, Mario Krumnow, Bernard Bäker, Jürgen Krimmling)

On-Line Traffic Modelling In Assen: The Sensor City **79**
(Klaas Friso, Kobus Zantema, Edwin Mein)

Reducing the Impact of Traffic Incidents using Capacity-Regulating Traffic Lights **89**
(Markus Rausch, Stefan Lämmer, Martin Treiber)

Simulation Study of a Traffic Light Assistant based on Vehicle-Infrastructure Communication 99
(Martin Treiber)

Vehicular Traffic Monitoring through VANETs: Simulation and Analysis in a Real Case Study 111
(Andrea Baiocchi, Chiara Colombaroni, Francesca Cuomo, Mario De Felice, Gaetano Fusco)

Air Traffic Management
(Andrea D'Ariano, Università degli Studi Roma Tre)

ITS Solutions for Air Cargo Revenue Management 123
(Tatjana Bolic, Lorenzo Castelli, Desirée Rigonat)

Air Traffic Optimization Models for Aircraft Delay and Travel Time Minimization in Terminal Control Areas 133
(Marcella Samà, Paolo D'Ariano, Andrea D'Ariano, Dario Pacciarelli)

A Mixed-Integer Optimal Control Approach for Aircraft Landing Model A.145
(Konstantin D. Palagachev, Matthias Rieck, Matthias Gerdts)

Water Traffic

Potential Fields in Maritime Anomaly Detection 145
(Ewa Osekowska, Stefan Axelsson, Bengt Carlsson)

Trajectory Optimisation for a Manoeuvre Guidance System in Inland Water Traffic 155
(Alexander Born, Iván Herrera-Pinzón)

Image Processing in Transportation
(Klaus-Peter Döge, TU Dresden)

Advantages and Limitations of a Dual Approach in Video-Based Traffic Data Acquisition 171
(Jan Grimm)

A Comparative Study of Shadow Models for Video-Based Traffic-State Analysis 181
(Klaus-Peter Döge)

Camera-Assisted Passenger Localization in Public Transport Vehicles 191
(Uwe Gosda, Richard Weber, Oliver Michler)

Johnson Criteria applied for Traffic Incident Detection Systems 201
(Johannes Traxler)

Vehicle Tracking using 3D Particle Filter in Tunnel Surveillance and Incident Detection 213
(Adrian Fazekas, Michael Bommes, Markus Oeser)

Localisation Technologies for Intelligent Transportation Systems
(Oliver Michler, TU Dresden and Ina Partzsch, Fraunhofer IVI)

Exploiting Vehicle Communication with Infrastructures for Accurate Positioning 223
(Gennaro N. Bifulco, Francesco Galante, Luigi Pariota, Spena Maria Russo)

Location Forwarding for Dense Urban Environments 233
(Alireza Ghods, Stefano Severi, Giuseppe Abreu)

Simulation of Wave Propagation for Radio and Positioning Planning inside Aircraft Cabins 243
(Julia Ringel, Samuel Klipphahn, Oliver Michler)

Floating Car Data
(Anita Graser, Austrian Institute of Technology GmbH and Matthias Körner, TU Dresden)

Network-Wide Application of Floating Car Data (FCD) particularly in Cities using Data Fusion with Measurement Data of Existing Stationary Traffic Detection 255
(Ralf Kohlen)

New Challenges in FCD Research 261
(Günter Kuhns, Elmar Brockfeld, Thorsten Neumann, Alexander Sohr, Louis Touko)

The Impact of Loop Detector Distance and Floating Car Data Penetration Rate on Queue Tail Warning 271
(Gerdien Klunder, Henk Taale, Serge Hoogendoorn)

Route Choice Identification and Selection from Sparse Floating Car Data Sets 281
(Gennaro Ciccarelli, Claudia Castaldi, Chiara Colombaroni, Gaetano Fusco)

Road-Bound Public Transport Management
(Jürgen Krimmling, TU Dresden)

Optimizing Public Transport Planning and Operations using Automatic Vehicle Location Data: the Dutch Example 291
(Niels van Oort, Daniel Sparing, Ties Brands, Rob M. P. Goverde)

Effects of Cooperative Traffic Signals on Tramway Operation 301
(Christian Gassel, Jürgen Krimmling)

Pre-signals for Bus Priority: Basic Guidelines for Implementation **311**
(S. Ilgin Guler, Monica Menendez)

Traffic Signal Preemption by Means of Digital Transmission Methods **323**
(Michael Preusker, Charlotte Gäbel, Stefan Löwe)

Railway Traffic Management
(Thomas Albrecht, TU Dresden and Francesco Corman, TU Delft)

Interfacing Conflict Resolution and Driver Advisory Systems in Railway Operations **333**
(Birgit Jaekel, Thomas Albrecht)

Kronecker Algebra based Modelling of Railway Operation **345**
(Mark Volcic, Johann Blieberger, Andreas Schöbel)

Role of Systems Engineering in Evaluation of ITS Systems – Example of The Train Dispatcher System (in Poland) **357**
(Grzegorz Karon, Jerzy Mikulski)

RTSE, a Multi-Component Closed-Loop Control Framework for Railway Networks **367**
(Raimond Wuest, Albert Steiner, Jonas Looser, Bernhard Seybold, Marco Laumanns, Juliane Dunkel, Daniel Huerlimann, Samuel Roos)

Railway Timetabling
(Thomas Albrecht, TU Dresden and Francesco Corman, TU Delft)

A Passenger Knock-On Delay Model for Timetable Optimisation **377**
(Peter Sels, Thijs Dewilde, Dirk Cattrysse, Pieter Vansteenwegen)

Capacity-Utilized Integration and Optimization of Rail Freight Train Paths into 24 Hours Timetables **389**
(Peter Großmann, Alexander Labinsky, Jens Opitz, Reyk Weiß)

The State-of-the-art Realization of Automatic Railway Timetable Computation **397**
(Michael Kümmling, Peter Großmann, Karl Nachtigall, Jens Opitz, Reyk Weiß)

Railway Dispatching
(Thomas Albrecht, TU Dresden and Francesco Corman, TU Delft)

Analysis of a Closed-Loop Control Framework in a Realistic Railway Traffic Environment **407**
(Egidio Quaglietta, Francesco Corman, Rob M. P. Goverde)

Boosting the Performance of a MILP Formulation for Railway Traffic Management in Complex Junctions 419
(Paola Pellegrini, Grégory Marlière, Joaquín Rodriguez)

Railway Traffic Control with Minimization of Passengers' Discomfort 429
(Francesco Corman, Federico Sabene, Dario Pacciarelli, Marcella Samà, Andrea D'Ariano)

Railway Data Analysis and Optimization
(Thomas Albrecht, TU Dresden and Francesco Corman, TU Delft)

Analyzing Railroad Congestion in a Dense Urban Network Through the Use of Road Traffic Network Fundamental Diagram Concept 439
(Pierre-Antoine Cuniasse, Christine Buisson, Joaquín Rodriguez, Emmanuel Teboul, David de Almeida)

Calibrating and Validating Train Dynamics Characteristics against Realisation Data 449
(Nikola Bešinović, Egidio Quaglietta, Rob M. P. Goverde)

Calibration of a Data-driven Railway Traffic Prediction Model 459
(Pavle Kecman, Rob M. P. Goverde)

Timetable Evaluation and Optimization under Consideration of the Stochastical Influence of the Dwell Times 471
(Anne Binder, Thomas Albrecht)

Traffic and Transit Assignment (Linked to COST Action TransITS)
(Francesco Viti, University of Luxembourg)

Combining Demand Management and Merge Control in an Equilibrium Network Model 483
(Francesco Viti, Wei Huang, Mike J. Smith)

Conflict Areas for Macroscopic Models in Dynamic Traffic Assignment 493
(Daniele Tiddi, Bojan Kostic, Guido Gentile)

Equilibrium and Day-to-Day Stability in Traffic Networks under ATIS 503
(Gennaro N. Bifulco, Giulio E. Cantarella, Fulvio Simonelli, Pietro Velonà)

Modeling Rerouting Phenomena through Dynamic Traffic Assignment in Rolling Horizon 513
(Guido Gentile, Rafał Kucharski, Lorenzo Meschini)

Travel Demand for a One-Way Vehicle Sharing System: a model of traffic assignment to a multimodal network with supply-demand equilibrium 523
(Fabien Leurent)

Table of Authors **535**

A Microscopic Simulation Approach for Optimization of Taxi Services

Michał Maciejewski[1, 2], **Kai Nagel**[2]
[1]Poznan University of Technology
[2]Berlin Institute of Technology

Abstract

This paper presents a simulation platform along with several on-line dispatching algorithms developed in order to optimize taxi services. First, the issue of simulation-based optimization of modern transport services, especially taxi services, is presented. Next, the proposed approach to microscopically simulate taxi services is explained, followed by a description of the on-line taxi dispatching algorithm framework and three selected dispatching strategies implemented within this framework. The next section presents the simulation scenario of Mielec that the strategies were tested on. Then, the simulation results obtained are analysed and the strategies compared. The paper ends with conclusions on the main properties and other possible applications of the proposed simulation approach, as well as on future plans concerning further improvements of the taxi dispatching algorithms.

Keywords: dynamic taxi dispatching, dynamic vehicle routing, demand-responsive transport, on-line optimization, simulation-based optimization, multi-agent simulation, MATSim, traffic flow simulation

1 Introduction

As a result of the development of Information and Communications Technology (ICT), transport services have become more intelligent, that is more flexible, demand-responsive, safe and energy/cost-efficient. This concerns both substantial improvements in the traditional means of transport, such as regular public transport or taxis, as well as the introduction of novel services, such as demand-responsive transport or personal rapid transit. However, the growing complexity of modern transport systems, besides all the benefits, increases the risk of failure due to the lack of precise design, implementation and testing. One way of dealing with this problem is the use of simulation tools that offer a wide spectrum of possibilities for validating transport service models (e.g. [Reg98; Bar07; Lia08]).

Concerning taxi services, such a simulation tool has to take into account the high dynamism of demand (since demand patterns change often daily and a significant number of orders are immediate ones, therefore it is hard to predict future demand accurately), the specificity of fleet management operations (due to the partial independence of taxi drivers, the dispatcher does not have a full control over them) and realistic traffic flow phenomena (as the urban traffic is highly dynamic). Therefore, the use of microscopic traffic flow simulation combined with microscopic travel demand models (e.g. activity-based) is crucial for accurate evaluation of the service under different circumstances, ranging from low to high load (e.g. large events, bad weather or public transport strikes). These issues, however, remain almost unexplored. To the best knowledge of the authors, traffic flow simulators have been applied only in Singapore [Lee04; Seo10]. Both approaches do not include realistic on-line demand generation; one cannot, for instance, model the impact of traffic or transport service availability on customer demand. Moreover, the systems were used only for small-scale problems, where a road network was of limited size and the number of customers was not high.

2 Microscopic Simulation of Taxi Services

As stated in Section 1, in the case of taxi services, a comprehensive approach should allow for running large-scale simulations with microscopically modelled demand, supply and traffic. Out of various simulation platforms considered, MATSim [Bal08] is arguably the one that is closest to meet all the requirements stated; see [Mac12] and references therein for additional justification. MATSim, however, does not provide taxi among the built-in means of transport, and therefore, has been extended and coupled with the DVRP Optimizer [Mac12] that is responsible for dispatching taxis. Whenever a new event occurs (such as a request submission, a vehicle departure/arrival) during the simulation in MATSim, the DVRP Optimizer reacts and adapts taxi drivers' schedules to the current state. Each taxi driver follows his/her schedule that consists of tasks of the following types:

- DriveTask – driving along a given route (the shortest path between two points).
- ServeTask – picking up a passenger at a given location.
- WaitTask – waiting at a given location for a new request.

A typical sequence of events associated with a single request is presented in Figure 1. A taxi customer calls the taxi service (request i; event $\mathbf{E_i^0}$ at time T_i^0) and waits until the taxi arrives. In response, the taxi dispatcher assigns the new request to one of taxis, according to a predefined algorithm (see Section 3). The selected taxi sets off for the customer ($\mathbf{E_i^1}$, T_i^1) and arrives at the pick-up location ($\mathbf{E_i^2}$, T_i^2). Then the customer is picked up and the taxi departs ($\mathbf{E_i^3}$, T_i^3). Finally, the taxi drops off the passenger at the destination location ($\mathbf{E_i^4}$, T_i^4). Since taxi demand and traffic are stochastic, times T_i^1, T_i^2, T_i^3 and T_i^4 are subject to change during simulation.

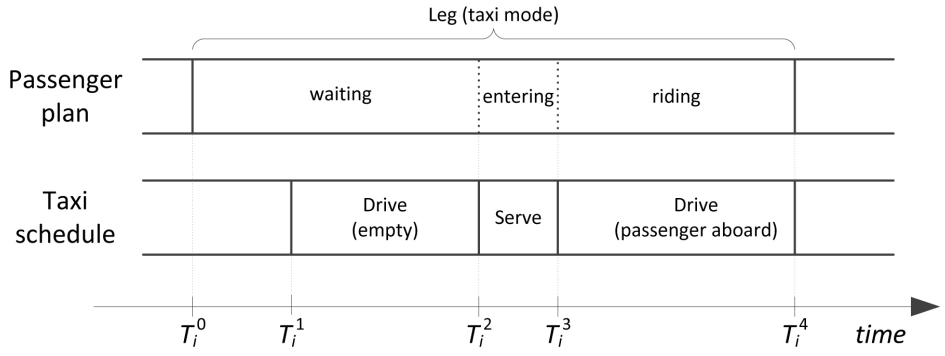


Figure 1: A planned taxi leg and the corresponding sequence of taxi tasks.

The traffic flow simulation in MATSim is based on queue approach, where each link is a FIFO queue and constitutes the basic network element. This is a simplification, as the FIFO approach does not allow vehicles passing each other. The advantage of this approach is that vehicles are processed computationally only when entering and leaving a link, allowing simulations to run at least a factor of 10 faster than those with dynamics on a link. Since we are mostly interested in the congested periods in an urban scenario, this is arguably an acceptable compromise as long as all vehicles that contribute to congestion have similar acceleration capabilities and maximum speed. For an approach to introduce vehicles of different capabilities see [Aga12].

The DVRP Optimizer operates on a directed graph where arcs are the shortest paths between a given pair of locations (i.e. links). As link travel times change over day, arcs are time dependent, i.e. their travel times and routes depend on departure time. By default, it is assumed that both link and arc travel times are calculated with the accuracy of 15 minutes. This is enough to reliably model the dynamics of traffic flow, and at the same time, limits the amount of shortest path searches (arc travel times and paths are cached for each time bin).

3 On-line Taxi Dispatching

Some studies propose the use of multi-agent approach to model (partial) independence of taxi drivers [Che11; Seo10; Als09], while other assume a fully centralised management [Lee04; Wan09]. All on-line taxi dispatching algorithms currently available in the DVRP Optimizer implement a certain pattern of collaboration between customers, taxi drivers and the dispatcher. Some algorithms are customizable and offer a choice between several patterns. A collaboration pattern is defined by the following set of properties:

- *destination knowledge* – the destination is known a priori if a customer informs the dispatcher about his/her destination.
- *request and taxi monitoring* – the dispatcher may monitor taxis and constantly update the timing of their schedules. Otherwise, taxi drivers notify the dispatcher only about switching between the *busy* and *idle* states.

- *requests reassignment* – already assigned requests can be dynamically reassigned between drivers. Request swapping is expected to be beneficial for both customers and drivers, and is usually coordinated by the dispatcher.

Each option implies some cooperation between interested parties. The first one involves additional customer-to-dispatcher communication, the second one imposes extra driver-to-dispatcher communication, while the last one requires real-time collaboration between drivers and the dispatcher.

Taxi fleets operate in a very dynamic environment, where demand, supply and traffic are stochastic and to some extend unknown. On-line taxi dispatching algorithms have to react to events representing changes in the system. In the simplest approach, commonly used by taxi companies, the dispatching procedure are performed only in response to submissions and completions of requests ($\mathbf{E_i^0}$ and $\mathbf{E_i^4}$, respectively). More sophisticated algorithms may be triggered also when taxis set off for, arrive at and depart from pick-up locations ($\mathbf{E_i^1}$, $\mathbf{E_i^2}$ and $\mathbf{E_i^3}$, respectively). Additionally, all taxicabs can be monitored on-line and the dispatching procedure may be executed for vehicles en-route if the expected arrival time changes.

In this paper, three dispatching strategies are analysed: *no-scheduling* that mimics the simplest approach, *re-scheduling* that monitors vehicles and responds to all possible events, and a modification of the latter that minimizes pick-up trip times instead of waiting times. All three strategies are based on the following assumptions:

- Dispatching procedures are fast enough to operate on a static snapshot of the current system state.
- Customers perform only immediate taxi calls and then wait for a taxi to come. To assure fairness, all taxi requests are scheduled based on the *first-come, first-served* policy.
- By default, *nearest taxi* means the nearest one in *time*. However, the first and third strategies may also use any measures of closeness in *space*.
- Although the second and third strategies may take advantage of the a priori destination knowledge [Mac13b], this aspect is beyond the scope of this paper – it is assumed that the destination remains unknown to the dispatcher until time T_i^3.

No-scheduling strategy (NOS) This strategy responds to $\mathbf{E_i^0}$ and $\mathbf{E_i^4}$ events in the following way:

- $\mathbf{E_i^0}$ – the nearest vehicle among the idle ones is dispatched to this request; if no vehicle is available at that time, the request is queued in the FIFO queue.
- $\mathbf{E_i^4}$ – the vehicle that has just completed request i is dispatched to the first request in the FIFO queue; if the queue is empty, the vehicle becomes idle.

Despite its simplicity, this strategy has several advantages over more elaborate ones. It has low demand for computational power. Additionally, it does not require travel times to be

known since it does not build schedules; one can even use straight-line distance to find the nearest idle taxi. Among the drawbacks is performance deterioration with the decreasing number of idle taxis — in an overloaded system, the first idle taxi may appear on the opposite side of a city.

Re-scheduling strategy (RES) This strategy extends the existing taxi schedules by appending a new request to the schedule of the nearest (in time) vehicle among all vehicles (both idle or busy), which requires the knowledge of travel times. This strategy monitors execution of requests and movement of vehicles. In response to some delays (or speed-ups), it updates the timelines of schedules and reassigns requests between taxis if the other one appears to be nearer. The strategy acts in the following way:

- $\mathbf{E_i^0}$ – request i is appended to the schedule of the *nearest* taxi
- $\mathbf{E_i^1}$ – $\mathbf{E_i^4}$ – if the vehicle serving request i is ahead of/behind time, full rescheduling is carried out (all planned requests are removed from schedules and scheduling is performed again according to the FCFS rule)
- link-to-link moves – if the vehicle is ahead of/behind time, the timing of its schedule is updated, while the assignments remain unchanged

This strategy considers all vehicles, both idle and busy, which increases the chances of finding better assignments. However, as destinations are unknown, the planning horizon is limited up to one pickup ahead (event $\mathbf{E_i^3}$), and therefore, vehicles with already one planned pickup cannot be considered when scheduling a new request (before arriving at the pick-up location). Frequent reassignments cause higher demand for computational power, compared to NOS.

Re-scheduling strategy for minimizing pick-up trip times (RES-PT) This strategy is a slightly modified RES, where the measure of closeness does not represent passenger's waiting times ($T_i^2 - T_i^0$) but pick-up trip times ($T_i^2 - T_i^1$), or other arbitrary pick-up trip distance measures. This modification shifts the preference from customers to taxi drivers. In an overloaded system, however, the reduction of pick-up trip times increases the system throughput, and hence, reduces the amount of time passengers wait for a taxi.

4 Test Scenario

In order to evaluate different strategies a simulation scenario was created in MATSim. The scenario represented a hypothetical private car traffic in the city of Mielec (south-eastern Poland, over 61,000 inhabitants) between 6:00 am and 8:00 pm, including both peak hours. The network consisted of 200 nodes and over 600 links and traffic was made up of over 56,000 private car trips. Detailed description of the model can be found in [Mac13a].

First, the simulation of Mielec was carried out for 20 regular iterations without taxis, which was enough for the relaxation process to converge. Next, a given fraction of intracity

private car trips were changed into taxi trips. Since such a conversion introduces extra traffic related to pick-up trips, the travel times increased. Additionally, the original private car paths may differ from the derived taxi drop-off paths, so the geographical layout of traffic may change. Therefore, before benchmarking each algorithm, five fully-functional (i.e. with taxis) warm-up iterations were carried out, which helped to obtain correct travel times. Finally, 20 benchmark iterations were executed.

The performance of the proposed optimization strategies was tested against different amounts of demand and supply. The taxi demand was modelled as $n = 406$, 636, 840, 1069, 1297, 1506 and 1719 requests, which corresponds to approximately 1, 1.5, 2, 2.5, 3, 3.5 and 4 per cent of private car trips converted into taxi ones. The spatio-temporal distribution of the taxi demand was identical to the regular traffic, as a result, the rush hours were the most challenging for taxi dispatching. On the supply side, the size of the fleet m was set to 25 and 50, and it did not change during a simulation period.

5 Simulation Results

Many different measures, all described in [Mac13a], were used to assess the performance of the proposed strategies, both from the taxi customers' and taxi company's points of view. In this paper, the following two measures are analysed: average passenger waiting time, $T_{\mathrm{W}} = \sum_{i \in N}(T_i^2 - T_i^0)/n$, and average pick-up trip time, $T_{\mathrm{P}} = \sum_{i \in N}(T_i^2 - T_i^1)/n$, where the former represents the customers' perspective and the latter defines the interest of the company. One may say that NOS and RES minimize T_{W} whereas RES-PT minimizes T_{P}. Figures 2 and 3 present the results obtained for the Mielec scenario and different n and m values. Separate curves were plotted for $m = 25$ (n/m between 16.24 and 68.76) and $m = 50$ (n/m between 8.12 and 34.38). All algorithms were run with time as the measure of taxi-to-request closeness. The results are averages over 20 benchmark iterations.

Figure 2 shows that neither of the strategies turned out superior for all demand-to-supply ratios. At low load, up to $n/m \approx 30$, NOS performs best while RES is slightly worse. At medium load, n/m between 30 and 55, RES is definitely the best performing strategy. However, at high load, above 55 requests per cab, RES-PT yields lowest T_{W}.

At first sight, the most curious is the advantage of NOS over RES at low load. One would expect that RES, having a broader choice set (idle and busy taxis) and using exactly the same travel time data, should outperform the former. That would happen in a scenario with uniformly distributed origin and destination locations, which is not true in the Mielec scenario, where people leave homes in the morning and return there in the evening. As a result, one can imagine a situation presented in Figure 4, where pick-up and drop-off locations are concentrated in two different parts of the city. Initially, request 1 is already submitted and cab 1 is idle. Soon request 2 will be submitted and cab 2 will become idle. In such a situation, NOS would assign request 1 to cab 1 and then request 2 to cab 2, while RES would do the opposite (assuming that cab 2 is *closer in time* to request 1 than cab 1). The decision made by RES is better from the perspective of request 1, which has priority

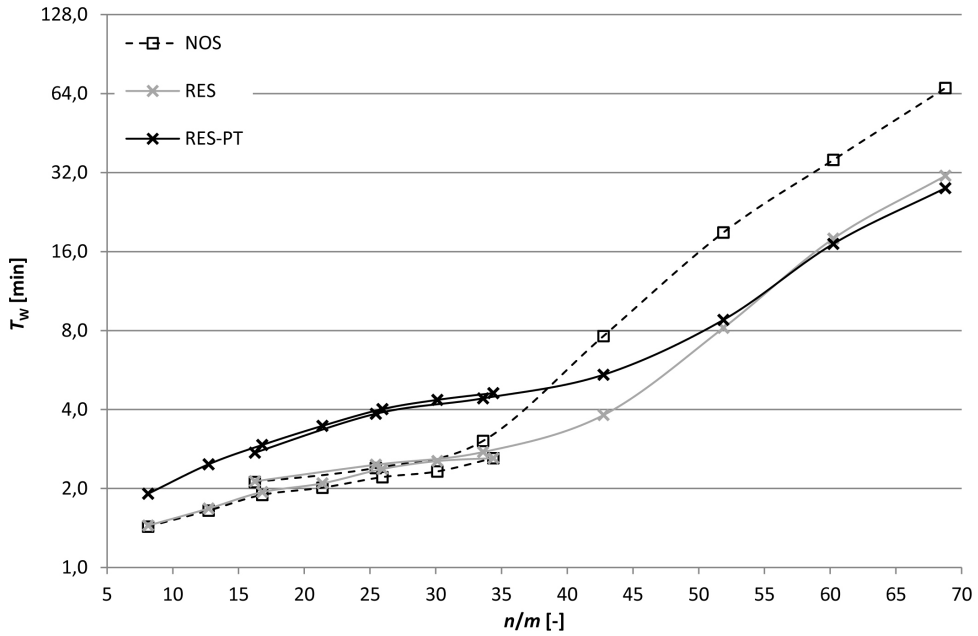


Figure 2: Average passenger waiting time T_{W} at different demand-supply ratios.

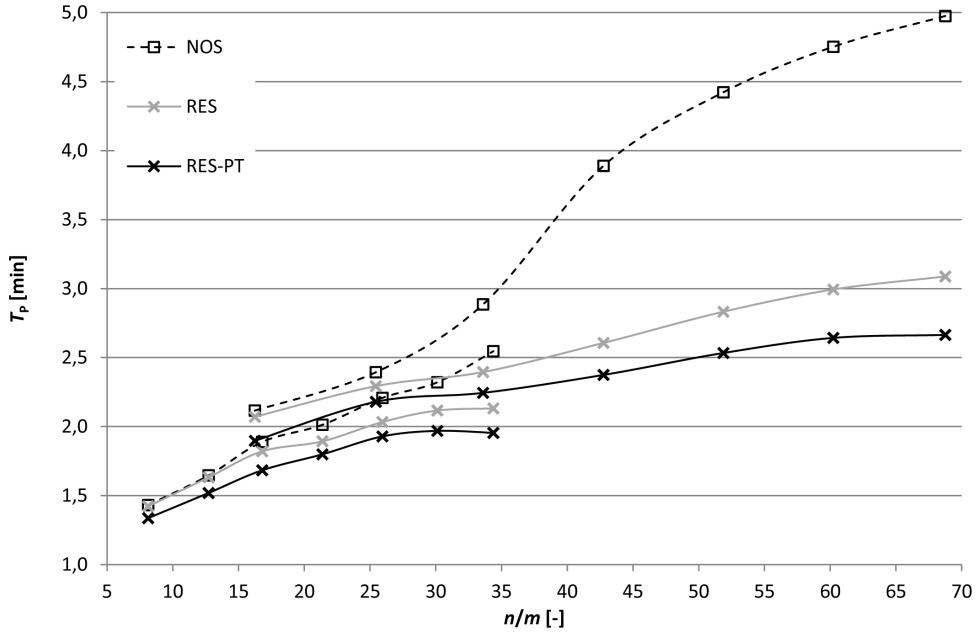


Figure 3: Average pick-up trip time T_{P} at different demand-supply ratios.

over request 2. However, the opposite would be better for the overall minimization of T_{W}, as it would increase slightly the waiting time of request 1 and, at the same time, reduce it significantly for request 2.

As expected (see Section 3), RES-PT outperforms RES at high load. By minimizing T_{P}, instead of T_{W}, taxis spend less time on pick-up rides, which, in turn, increases the system

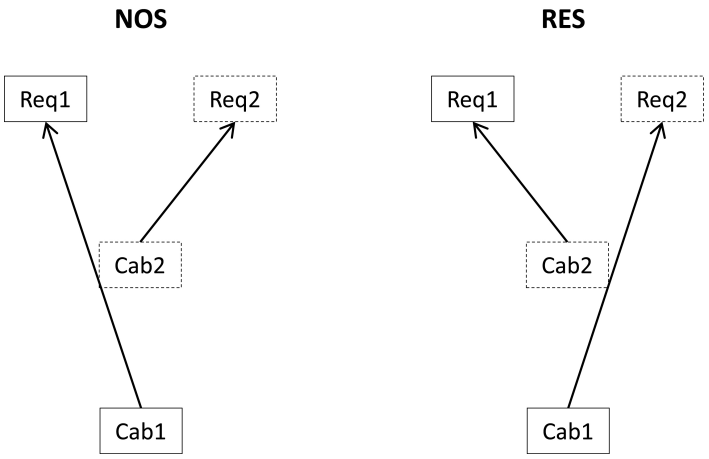


Figure 4: A comparison of the operation of NOS and RES.

throughput. Eventually, higher throughput means lower T_{W}.

Looking at Figure 3, the results there seem very plausible. RES-PT gives the lowest values since the main aim of this strategy is the minimization of the pick-up trip times. RES is better than NOS, as the possibility of choosing a busy taxi reduces the pick-up trip times — not only does it help in minimization of T_i^2, but also it allows for higher T_i^1, which eventually results in lower T_{P}. At low load, where almost all taxis are idle, NOS and RES perform similarly. This changes as demand rises and the number of idle taxis drops, which narrows the choice of taxis in NOS.

There is an interesting relation between 25- and 50-taxi series in Figure 3. They are not adjacent and we obtain higher values for smaller fleets. This is due to the fact that the average distance to the closest taxi drops with the growing number of taxis. The same pattern occurs in Figure 2 for NOS, and partially for RES, however, not for the RES-PT series. This is caused by the fact that minimization of T_{W} (NOS and RES) requires T_{P} to be small as well. Additionally, NOS results in the equality $T_{\mathrm{W}} = T_{\mathrm{P}}$ as long as the system is not overloaded (i.e. there is always at least one taxi at the dispatcher's disposal; in the Mielec scenario, this is true up to $n/m \approx 30$). On the other hand, minimization of T_{P} (RES-PT) does not require low T_{W}. Moreover, it works better if the choice of awaiting requests is large, which relates to higher values of T_{W}. For example, one can imagine a strategy that waits until the end of a day (i.e. until all requests are submitted) and then builds routes for taxis. This strategy would provide very low T_{P} but at the cost of unacceptably high T_{W}.

Last but not least, all strategies fulfil the real-time execution criteria; in the case of the Mielec scenario, the total computation time spent on optimization varies between 0.5 and 4 seconds (depending on the strategy, demand and supply), whereas traffic flow simulation takes less then 1 second (on the Intel Core i7-2600K processor). Of course, running bigger scenarios, with larger and more detailed networks, higher supply and demand, will result in longer computation times.

6 Conclusions

The developed simulation system combining MATSim and the DVRP Optimizer proved to be useful for realistic simulation of taxi services. The high level of detail used for describing demand, supply and traffic allows for modelling precisely collaboration between the main actors, that is customers, taxi drivers and the dispatcher, all embedded into a larger transport system of a city. This collaboration takes advantage of modern ICT solutions that enable the dispatcher to smoothly coordinate the fleet, including (re-)assignments of requests to taxis. Although the Mielec scenario is not a fully real-life scenario (some data were generated artificially), there is an ongoing work on simulation of taxi dispatching for Berlin and Poznan (Poland's fifth largest city) that is both microscopic and large in its scale. Moreover, after some adaptations, the software may be used for a broad spectrum of vehicle routing and scheduling models (e.g. emergency services, demand-responsive and shared transport, or commercial fleet operations) in order to facilitate development of efficient ITS systems.

The detailed analysis of the simulation results gives us insight on the characteristics and performance of various implemented dispatching strategies, out of which three have been described in this paper. The outcomes obtained show that neither of them is best, and therefore, there is a need for a kind of 'super strategy' that would combine different qualities of the original ones. Compared to RES, the new strategy should apply a broader perspective when assigning requests to taxis. In particular, it should consider all awaiting requests (at high load), anticipate future ones (at both low and high load), and finally, focus not only on the first queued request but on the whole system's performance.

References

[Aga12] A. AGARWAL, M. ZILSKE, K. RAMACHANDRA RAO, and K. NAGEL: *Person-Based Dynamic Traffic Assignment for Mixed Traffic Conditions*. VSP Working Paper 12-11. TU Berlin, Transport Systems Planning and Transport Telematics, 2012. URL: www.vsp.tu-berlin.de/publications.

[Als09] A. ALSHAMSI, S. ABDALLAH, and I. RAHWAN: "Multiagent self-organization for a taxi dispatch system". In: *8th International Conference on Autonomous Agents and Multiagent Systems*. 2009, pp. 21–28.

[Bal08] M. BALMER, K. MEISTER, M. RIESER, K. NAGEL, and K. AXHAUSEN: *Agent-based simulation of travel demand: Structure and computational performance of MATSim-T*. Tech. rep. ETH, Eidgenössische Technische Hochschule Zürich, IVT Institut für Verkehrsplanung und Transportsysteme, 2008.

[Bar07] J. BARCELO, H. GRZYBOWSKA, and S. PARDO: "Vehicle routing and scheduling models, simulation and city logistics". In: *Dynamic Fleet Management: Concepts, Systems, Algorithms & Case Studies*. Ed. by V. ZEIMPEKIS, C. D. TARANTILIS, G. M. GIAGLIS, and I. MINIS. NY: Springer, 2007, pp. 163–195.

[Che11] S. CHENG and T. NGUYEN: "Taxisim: A multiagent simulation platform for evaluating taxi fleet operations". In: *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Volume 02*. IEEE Computer Society. 2011, pp. 14–21.

[Lee04] D. LEE, H. WANG, R. CHEU, and S. TEO: "Taxi dispatch system based on current demands and real-time traffic conditions". In: *Transportation Research Record: Journal of the Transportation Research Board* 1882.-1 (2004), pp. 193–200.

[Lia08] T.-Y. LIAO, T.-Y. HU, and D.-J. CHEN: *Object-Oriented Evaluation Framework for Dynamic Vehicle Routing Problems Under Real-Time Information*. Annual Meeting Preprint 08-2222. Washington D.C.: Transportation Research Board, 2008.

[Mac12] M. MACIEJEWSKI and K. NAGEL: "Towards multi-agent simulation of the dynamic vehicle routing problem in MATSim". In: *Parallel Processing and Applied Mathematics (PPAM), Revised Selected Papers, Part II*. Ed. by R. WYRZYKOWSKI ET AL. Lecture Notes in Computer Science. Springer, 2012, pp. 551–560. DOI: 10.1007/978-3-642-31500-8_57.

[Mac13a] M. MACIEJEWSKI and K. NAGEL: *Simulation and dynamic optimization of taxi services in MATSim*. VSP Working Paper 13-05. TU Berlin, Transport Systems Planning and Transport Telematics, 2013. URL: www.vsp.tu-berlin.de/publications.

[Mac13b] M. MACIEJEWSKI and K. NAGEL: *The influence of multi-agent cooperation on the efficiency of taxi dispatching*. VSP Working Paper 13-10. TU Berlin, Transport Systems Planning and Transport Telematics, 2013. URL: www.vsp.tu-berlin.de/publications.

[Reg98] A. REGAN, H. MAHMASSANI, and P. JAILLET: "Evaluation of Dynamic Fleet Management Systems: Simulation Framework". In: *Transportation Research Record* 1645 (1998), pp. 176–184.

[Seo10] K. SEOW, N. DANG, and D. LEE: "A collaborative multiagent taxi-dispatch system". In: *Automation Science and Engineering, IEEE Transactions on* 7.3 (2010), pp. 607–616.

[Wan09] H. WANG, D. LEE, and R. CHEU: "PDPTW based taxi dispatch modeling for booking service". In: *Natural Computation, 2009. ICNC'09. Fifth International Conference on*. Vol. 1. IEEE. 2009, pp. 242–247.

Corresponding author: Michał Maciejewski, Poznan University of Technology, Faculty of Machines and Transportation, Institute of Machines and Motor Vehicles, ul. Piotrowo 3, 60-695 Poznan, Poland, phone: +48 61 665 5957, e-mail: michal.maciejewski@put.poznan.pl;
Berlin Institute of Technology, Institute for Land and Sea Transport Systems, Transport Systems Planning (VSP), Salzufer 17-19 Sekr. SG12, D-10587 Berlin, Germany, phone: +49 30 314 23308, e-mail: maciejewski@vsp.tu-berlin.de

Challenges for Better Understanding and Simulating Urban Traffic - the Zurich Experience

Christian Heimgartner[1], Monica Menendez[2]

[1] City of Zurich

[2] Swiss Federal Institute of Technology Zurich

Abstract

The establishment of Intelligent Transport Systems and Services (ITS) seems to be nowadays an indispensable step for the promotion of more sustainable transport systems, especially in urban networks. In order to evaluate these strategies, often times we heavily depend on traffic models and simulations. Unfortunately, the results from all these models are only as good as the models themselves. Our ability to emulate reality is constrained by the underlying formulations of the used model, as well as our selection of input parameters. These elements are crucial to the quality and reliability of results. However, their importance is often underestimated, and results are taken from granted without a proper study of the simulated scenarios, and their similarities (or lack of) with the real networks been analysed.

This paper builds on the experience of the City of Zurich on modelling urban traffic. The goal is to highlight the complexities associated with the process, the different steps taken throughout, and the existing need for further research in some specific areas. Special attention is given to the integration of macroscopic demand models with microscopic traffic models; the availability of new data sources; the interactions between different transport modes; and the need for a better understanding of local conditions. We believe that to reach high quality performance when simulating urban traffic, it is important to improve our observation/monitoring methods, increase the accuracy of our modelling tools, and improve our calibration procedures.

Keywords: micro-simulation, calibration, urban traffic, transport models, ITS

1 Introduction

Sustainability has become a main issue in many cities around the world. The increasing population density and transport demand, and the resulting externalities (e.g. pollution, noise, energy consumption), have made the promotion of sustainable development a big

challenge for many city authorities. One promising approach for facing this challenge is the establishment of Intelligent Transport Systems and Services (ITS). They aim at improving the transportation system through the use of information and/or adaptive and self-management capabilities that respond to transport conditions on real time.

Unfortunately, due to the novelty of these systems and services, often times they have not been fully tested on the specific conditions for which they will be implemented. Nonetheless, most city authorities demand reliable facts before making ITS investment decisions. Some of these facts can be compiled from best practices and related tools [2DE11]. Experiences from other authorities with same or similar ITS measures have to be interpreted with caution, paying attention to the circumstances under which they were implemented. Such circumstances may differ significantly across locations and time periods. In those cases, as well as in the cases where the ITS measures are simply too new and have not been implemented before, transport modelling is a suitable tool (sometimes the only one) to assist decision makers.

Transport models and simulations allow us to analyse the effects of different ITS measures both for the actual situation, and for different forecasted scenarios. Thus, city authorities can perform ex-ante evaluations of ITS. Microscopic traffic models are particularly useful for this, as they model vehicular traffic (sometimes interacting with other modes) at an operational level. Thereby, accurate estimation of traffic related indicators at a disaggregated level (e.g. queue length, still stand time, number of stops) is possible. This is important, as the interactions among vehicles and with traffic control facilities are often crucial to the results of ITS measures. Moreover, due to its disaggregated nature, such models can be linked to other tools to estimate other impacts and externalities, e.g. pollution, noise, energy consumption (see [Tec12], [Ede12] and [ICT13] for examples).

Unfortunately, the results from all these models are only as good as the models themselves. Our ability to emulate reality is constrained by the underlying model formulations, and our selection of input parameters. These elements are crucial to the quality and reliability of results. However, their importance is often underestimated, and results are taken from granted without a proper study of the simulated scenarios, and their similarities (or lack of) with the real networks been analysed.

2 The Zurich approach

As the City of Zurich follows the vision of sustainability and its associated strategies (e.g. [Zur11]), there is a huge motivation to promote related ITS measures even though the promotion of sustainable measures is a rather complex issue. The road network in Zurich corresponds to a non-uniform layout, with mainly narrow streets and intersections close to each other. This network is shared by many public transport lines (i.e. buses and trams) besides motorised individual traffic and soft modes (i.e. pedestrians and bicycles). In addition, most of the traffic signals are actuated to provide full priority to public transport.

In order to look for sustainable strategies to address the complexity associated with those characteristics while accommodating the increasing demand, the Division of Transport in the

City of Zurich (DTZ) established a multimodal micro-simulation model. The model is linked to the traffic control programs of Zurich, and uses all common state-of-the-art model resources. In its current state it covers the city centre, an area of 2.6 km². It is mainly focussed on public transport and motorised individual traffic, and is intended as a permanent maintained microscopic model. The goal is for it to be the main tool for testing ITS strategies as well as other potential measures to improve transport systems in the city. Figure 1 shows the overall framework for the model and its interfaces. More details are provided below.

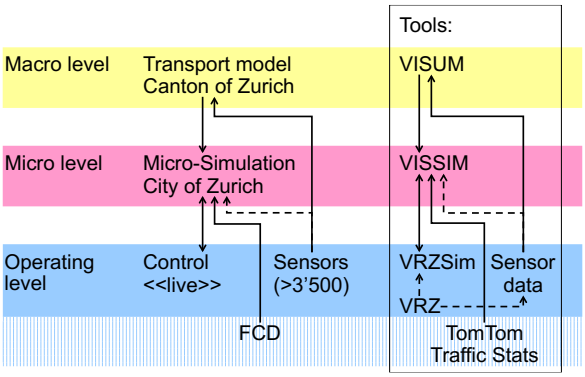


Figure 1: Framework for modelling transport in the City of Zurich. Source: [Hei12].

On the one side, the VISSIM (traffic simulation software, PTV Group, Karlsruhe) micro-simulation model is based on the transport demand model of the Canton of Zurich [Zur10]. As this demand model was established with VISUM (transport assignment modelling software) and VISEVA (transport demand modelling software), the interface from VISUM to VISSIM offers the possibility to connect the network elements and demand data (Figure 2). Demand from VISUM is exported directly into VISSIM in the form of origin-destination tables and assigned routes. To implement this connection the demand model network was modified to reach a level of disaggregation consistent with the micro-simulation model requirements (for details see [Fro12], [Hei12]). The overall tool and interface is maintained in order to ensure that the so obtained consistency between the microscopic and macroscopic models is kept in case of updates to the demand model. This involves not only technical, but also organizational challenges, as the two models belong to different authorities.

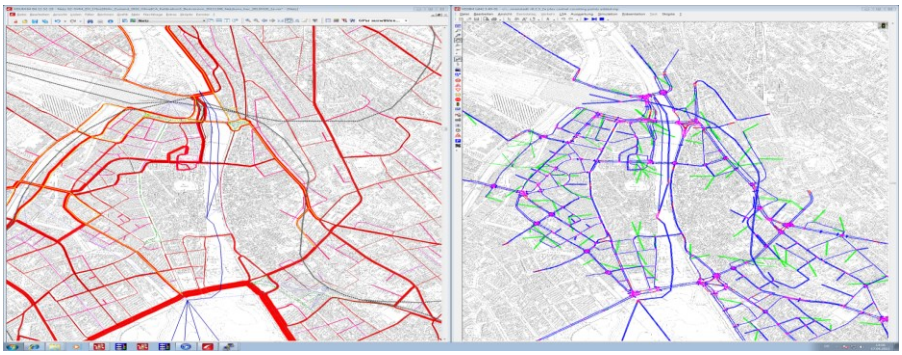


Figure 2: Underlying macroscopic demand model (left), and traffic micro-simulation model (right). Source: [Hei12] and [Zur10].

On the other side, the VISSIM micro-simulation model is linked with the operational level control schemes in the city of Zurich. In that regard, the core element is the interface that links the micro-simulation model with the signal plans. As Zurich's traffic control system

(VRZ) is not a standard product, it was necessary to develop a special interface. Given that the control system is clocked on a real time based process flow, and VISSIM has its own and variable process speed, it was not possible to create a direct link. The solution was found in VRZSim, a simulation that emulates the traffic control system, linked to VISSIM. Since no standard interface existed for that, a new one was developed.

A second link to the operational level concerns the use of sensor based traffic counts. These counts were used for the calibration of both the macroscopic demand model, and the microscopic traffic model. Moreover, through this linkage the counts can also be used for validating purposes or as quantitative bases for the establishment of certain specific traffic volume situations.

A third and final link to the operational level concerns the use of Floating Car Data (FCD) from TomTom Traffic Stats [Tom11]. That FCD was used as the foundation for an overall concept of model calibration, which included not only the calibration with traffic counts, but also with speeds. More details are provided in Section 3.

It is worth noticing that the traffic simulation model has already been used successfully for project works for the City of Zurich and a case study [Hei12], as well as some research projects [Ort12][Sch13][Ge13a], taking into account the challenges figured out later on in this paper.

3 Ways to high quality performance

VISSIM, the software used by the DTZ, is a microscopic, time step and behaviour-based simulation model. It is a well-known commercial software with many applications and high potential. It has proved to be a valid tool to model specific links and intersections, as well as large networks. Nonetheless, before any results can be expected from it, it must be properly calibrated as is the case of any other simulator.

In the case of the Zurich network, a detailed calibration process was undertaken. The goal was to calibrate for both flows and speeds, in order to ensure that the model replicated reality as much as possible. However, given the complexity of the network (see details in Section 2) and the process itself, a manual calibration was not feasible. VISSIM currently includes 192 parameters, and evidently, the calibration of all of them is simply impossible. A more feasible solution is the calibration of just a few parameters, the ones that the model is more sensitive to. The selection of those parameters is rather important. A calibration performed with an incomplete (or wrong) set of parameters may lead to multiple issues, including but not limited to, model imprecisions, and unrealistic values for the calibrated parameters. Regrettably, despite the importance of the process, there is usually no standard procedure for it [Mul11]. A sensitivity analysis (SA) is sometimes recommended as an intermediate step, especially for computationally expensive models. The SA can provide both quantitative and qualitative information regarding the effects of the different model input parameters (and their variations) on the simulation results [Sal08]. This can be helpful for identifying the most important parameters (i.e. those that should be included in the calibration process). Unfortunately, to the authors' knowledge, there are very few examples

of the application of SA in the calibration of microscopic traffic models [Ge13a]. The huge quantity of parameters contained in commercial simulators, combined with the fact that many of them are unobservable in the field and/or hard to measure, can make the SA also rather difficult. In addition, for the case of the Zurich network, given not only the size and complexity of the network, but the interface with the traffic signals control scheme (VRZSim), running simulations is very time consuming (around 30 minutes for a 1-hour simulation). This further complicates the process. Below is a brief explanation of the methodology used.

The sensitivity analysis was divided into three steps. The first step was done based on Zurich inner city's traffic patterns and characteristics, model intended uses and available data. 148 relevant parameters were chosen among the 192 (e.g. parameters related to bicycle lanes were discarded as bicycles were not included in the model). The second step was done based on the existing literature, common sense, and the authors' experience. Values for 55 parameters were extracted from the macroscopic demand model; default values were used for 79 parameters regarded as not very influential or not highly variable; and 14 parameters were selected for a more in-depth analysis. For the third step, those 14 parameters were evaluated according to their influence on the model outputs. Given the lack of standard procedure to conduct the SA of computationally expensive microscopic traffic models, for this we developed a new method: the quasi Optimized Trajectories Elementary Effects (quasi-OTEE), based on a modification of a well-known SA method (for more details see [Ge13b] and [Ge13c]). The proposed approach was used as a preliminary screening tool to find the parameters with the biggest influence on travel time (a proxy for speeds in the system). An automatic SA tool was developed based on this approach and implemented through programming with the VISSIM COM interface and MATLAB. For the travel time measurements, data was collected for one hour, after a 15 minutes warm-up period, in 20 road sections. Notice that these road sections were rather large relative to the size of the network, and the sum of them covered almost all the streets in the network.. As the lengths of the 20 sections were different, we used the *Travel Time per Meter Traveled* (TTMR, calculated as travel time divided by the length of the corresponding section) (see [Men12] for more details), and compared with TomTom Traffic Stats. The obtained sensitivity indexes (i.e. μ^* (absolute mean), and σ (standard deviation) of the elementary effects) of each parameter are plotted in Figure 3. Notice that the values for the sensitivity indexes are context dependent and their scale could be totally different in other studies.

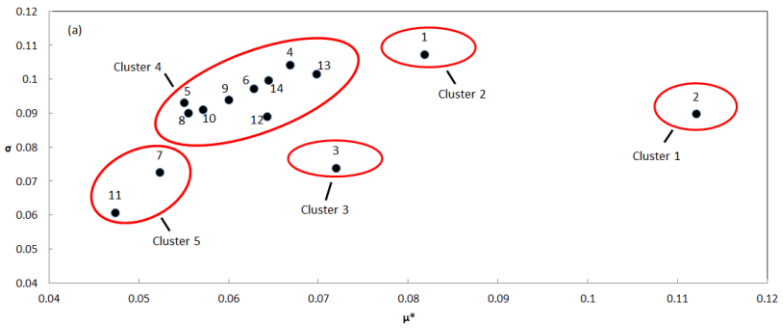


Figure 3: Plots of μ^* versus σ of the elementary effects with respect to TTMR for the 14 parameters. Source: [Ge13c].

In order to systematically separate the parameters into different groups according to their importance, we used K-Means Clustering [Mac67]. As can be seen, parameters in cluster 5 have the lowest μ^* and σ, therefore they are deemed to be the least influential ones (i.e. they are almost independent of the other parameters and the model is rarely sensitive to their variations). The parameter in cluster 1 has the highest μ^* indicating that it is the most influential parameter. In addition, its high σ indicates that it has strong non-linear effects and/or interactions with other parameters. Same is true for the parameters in clusters 2 and 3. As for the parameters belonging to Cluster 4, further analysis revealed that they all have relatively low μ (mean of the elementary effects) but high μ^* and σ. This indicates that these parameters will have both positive and negative effects depending on the adopted values of other parameters. Therefore, in order to avoid the Type II error (i.e. considering an important parameter as non-important), we expanded the set of most influential parameters by including Parameters 4, 13 and 14 (i.e. the parameters with the highest μ^* and σ in Cluster 4). To sum up, Parameters 1, 2, 3 (the Average Standstill Distance, the Additive Part of Desired Safety Distance, and the Multiplicative Part of Desired Safety Distance, from the car-following model), Parameter 4 (the Maximum (Own) Deceleration from the lane-changing model), and Parameters 13 and 14 (the Lane Changing Distance, and the Emergency Stop distance, from the lane model) are the most influential parameters in this case study. For a full definition of all the parameters and more details on the analysis see [Ge13c]. Notice that these results differ slightly from previous evaluations carried out in older versions of the network (i.e. the network has been modified to improve the layout, etc.). Although the majority of the most and least important parameters remain the same, their ranking has changed slightly.

Figure 4 shows the resulting flows and travel times for the 20 measurement sections after a preliminary calibration was carried out based on the SA results described above. The graphs reveal the significant differences in travel times between the simulation and the TomTom Traffic Stats. The VISSIM Zurich model consistently over-predicted speeds for all possible values of the analysed parameters. As a result, and although calibration of flows was successfully carried out, the calibration of speeds could not be finished. It is pending on further improvements to the network (more details are discussed in Section 4). Once the network is fully ready, such calibration can also be done automatically by modifying the values of the selected parameters. For that we developed a tool based on the Pattern Search algorithm [Kol03]. This software was implemented via VISSIM COM interface programming, and it can calibrate the model fast and efficiently. Caution must be exercised during that part of the calibration process, in order for the flows to continue to match reality even when parameters are changed to match speeds.

4 Facing the challenge

Given that the differences between simulation results and empirical values were unexpectedly large, some questions about the models/process arose. Through these questions we tried to gain a deeper understanding on those discrepancies [Hei12].

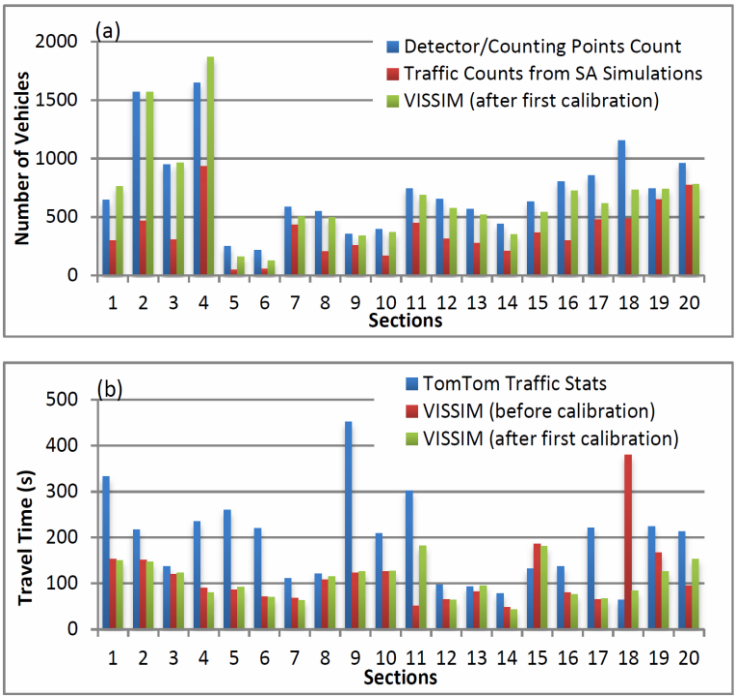


Figure 4: (a) 1 hour traffic counts. (b) Travel times. Source: [Men12].

1. *Are the TomTom Traffic Stats and other traffic counts inaccurate?*

We performed an additional validation of the FCD from TomTom Traffic Stats by collecting the data ourselves for the nine most problematic sections. The experiments showed that the FCD was plausible. However, its real characteristics could not be identified in a concluding manner due to the confidentiality used by the data suppliers.

2. *Is the quality/performance of the VISSIM model not high/good enough?*

We noted that small network errors in VISSIM could have a big impact on the traffic flows. For example, in one case an extra lane for turning was too short, in another case there was a one-lane connector between two links with two lanes each. The result in both cases was not enough traffic flow downstream. These mistakes were initially made in the macroscopic network, however their effects lingered through the microscopic model as well. Another issue was the absence of uncontrolled pedestrian crossings in VISSIM, which do have a significant impact on traffic flows and travel times in the real network. In order to address these issues, all the identified network errors were corrected, and additional pedestrian crossings were added into VISISM to emulate uncontrolled crossings in Zurich. Unfortunately, it is unfeasible to estimate the influence of any left-over errors until they are fully identified and corrected. Hence, the network continues to have minor errors that are detected and fixed on an on-going base.

3. *Is the quality/performance of the refined VISUM model not high/good enough?*

As the VISUM model was calibrated on the basis of traffic counts located on certain links, it can be assumed that it is most accurate concerning these links. The turning values in intersections, however, are not so accurate (as they were not specifically calibrated). In one case, for example, the wrong turning value caused a non-realistic blockage of the lanes going

straight within the VISSIM simulator. Efforts were made to correct all turning values based on real data/experience. However, although we have visually checked the presence of congestion in VISUM and the resulting VISSIM network, calibration only based on flows cannot guarantee that the proper density/speed is achieved.

4. *Does the connection between VISUM and VISSIM have systematic problems?*

The method to combine a macroscopic demand with a micro-simulation model can be seen in an ambivalent way. On the one hand, it is actually the only way to implement the proper demand patterns in a big micro-simulation network. On the other hand, the demand model does not take into account all the operational aspects that are important at the microscopic level. For example, in macroscopic demand models car density is not considered for the calibration process. However, it is highly relevant for the microscopic traffic models (e.g. same traffic counts might correspond to different density levels in peak and off-peak hours respectively). As a result, especially in peak hours, it is possible to underestimate traffic flow levels in certain non-calibrated areas of the network, if density (or occupancy of detector loops) is underestimated in other (even calibrated) areas. To address these issues, we recommend to look at both, flows and densities when integrating macroscopic demand models with microscopic traffic models.

5. *Was something else overlooked?*

In order to address other possible causes of discrepancies after all corrections described above were made, we performed multiple experiments with the VISSIM network by changing different aspects of it. For example, we ran a number of simulations with variations across random seeds, demand levels, warm up times, and desired speed distributions [Men12]. At the end, traffic counts were fitted more accurately, but travel speeds remained significantly overestimated (see Figure 4). The differences were still so high that the automatic calibration process was not applicable. The surveyed values could not be fitted with a reasonable variation of the input parameters. Given that significantly higher demand levels, combined with lower desired speeds did not lead to better results, it became clear that the solution was not trivial at all.

Evidently, given the complexity of the issues at hand, and the importance of their solution for the proper modelling and simulation of ITS, this is quite a big challenge and task for the research community. As it has become evident with this project, decision makers today cannot get completely reliable facts from ex-ante traffic network oriented analysis. The results of any traffic related impact analysis must be interpreted with regard to the above described aspects, especially because the current general understanding of urban traffic flows is still very limited. Therefore, and due to the fact that urban traffic flows are often unstable, over saturated, and lead to costly externalities, the above outlined challenge has to be faced to stay consistent with a strategy of sustainability. Hence, the motivation to promote related research activities is big for city authorities. We now understand better where to focus our future efforts.

4.1 Increasing accuracy in observing/monitoring urban traffic

First we must gain new insights through observing/monitoring urban traffic using both traditional and state-of-the-art methods. While the supply, including road infrastructure and urban traffic management is nowadays easily observed, ITS offer new potential for monitoring other aspects of urban traffic even in real time. For example, with regard to the transport demand, in addition to traffic counts and loop detector occupancies, FCD based on GPS or Bluetooth technology offers new possibilities to identify real trajectories of vehicles and persons, and therewith related traffic flows. This could be helpful for identifying which demand patterns refer to which operational situation, and how to accurately define them (e.g. origins, destinations, turning values, and routes including the 'last mile' per mode); as well as understanding the ability of drivers to adapt to new planned or unplanned situations (e.g. changes in routes to avoid congestion). The combination of new and traditional sources of data could establish a good basis for demand modelling and its calibration. Note, however, that FCD in most cases might be obtained from companies that are not willing to provide full transparency on the data collection/processing methodology, so the confidence in such data is an important issue. Regarding travel and driving behaviour there is the need for further empirical analysis (e.g. video based) and visual on-site observations. Specific attention should be paid to: unregulated pedestrian crossings and its effects on traffic; cooperative behaviour in intersections/networks in case of (over)saturation; speeds of vehicles at, and approaching intersections specially compared to those in the simulation; illegal driving manoeuvres; conflicts and other interactions between private vehicles and other modes (e.g. bicycles, buses, trams, pedestrians); on street parking and its negative effects on the road capacity and the moving traffic; and lane changing behaviour. Notice that some of these aspects do not vary much as a function of the location and could be applied similarly to analogous situations. On the other hand, local aspects are explicitly dependent on the specific local situation, and their value for the proper modelling of real scenarios should not be underestimated.

4.2 Increasing accuracy in modelling urban traffic

Based on an accurate observation of the network, the second step is to improve the performance of urban traffic models. Nowadays, micro-simulation models typically need information about origins and destinations of trips. If prediction of scenarios and their related effects on the demand is an issue to investigate, then the integration of a demand model is inevitable. To do this it could be suitable to integrate a microscopic agent based approach (e.g. MATSim [Ins13]). However, as of today, the computational requirements of this type of models is very big for microscopic and operation oriented networks. Hence, there is still a need to aggregate the microscopic operational aspects of urban traffic to a macroscopic level. As software tools offer more applications to exchange data between different levels of aggregation this process should become at least feasible.

Microscopic simulators can typically model the network and traffic rules rather well for both private and public transport. In regards to soft modes, the situation is different. Even though the modelling of pedestrians has rapidly developed in the last years, still major

improvements are needed for modelling cycling. To model travel and driving behaviour, approaches like the car following model are well established. Caution must be exercised, however, given that even this type of models might have some weaknesses, especially in (over)saturated networks and intersections, where behaviour becomes more cooperative, and drivers more adaptive. Research on using alternative approaches as for example the social force model (see [Hua12]) has to be strengthened.

At the macroscopic level, some improvements were achieved especially in the last year to include many operational elements (e.g. signal plans, detectors). Nonetheless, the accurate modelling of these elements to better replicate urban traffic conditions (e.g. influence of pedestrian crossings on speeds, accurate aggregation of actuated traffic control) remains an issue for further improvement. In terms of demand modelling, more research is required to evaluate how additional insights, especially gained from FCD, could assist in obtaining a macro model that fits better from an operational point of view (e.g. calibration of trip distribution, choice of trip start time). In that regard, we must proceed with caution when using traffic counts from congested areas for calibration, as this might lead to an underestimation of car density and the related congestion effects. In addition, it would be worthwhile to check the possibility for estimating demand patterns based on real time and predicted data, possibly in combination with short term prediction tools (e.g. OPTIMA real-time traffic forecast and traffic management decision support tool, PTV Group, Karlsruhe). Searching the most suitable dynamic oriented assignment methods with regard to the points addressed above is also an important task. State-of-the-art static traffic assignment may lead to link and node loads that overestimate traffic capacities (see [Bue04]). This is related to the fact that the whole demand goes through the network during the regarded time period (e.g. peak hour), without taking into account further effects of congestion on link and node loads (e.g. by a demand transfer to an earlier or later time period). One approach could be the increased integration of assignment at the micro level. Taking into account soft modes, related interactions, and intermodality, poses additional challenges for the above mentioned tasks. Nevertheless, modelling that accurately, as well as their effects on capacities and speeds could lead to additional insights.

4.3 Increasing accuracy in calibrating urban traffic models

On the basis of the above proposed investigations the calibration process has to be emphasized more. In addition to the calibration method described in Section 3, we believe that an in-depth SA examining different demand elements (e.g. number of trips between origins and destinations, number of trips per route, turning values) might be highly useful. This could definitely help identify which aspects of the overall modelling approach on the micro and the macro level are really significant.

It is our opinion that the aspects described above have to be addressed and put together like a patchwork/puzzle. It is an iterative process: to establish high quality urban traffic simulators, we must improve our monitoring schemes, increase the accuracy of our modelling tools, and develop our calibration procedures. Although the challenge is substantial and the increased weight of soft modes does not make it easier, this is the best

path to establish reliable decision support tools for ITS investments. If we intend to promote ITS as a measure of sustainability, these challenges have to be faced.

References

[2DE11] 2DECIDE CONSORTIUM: *2DECIDE, Toolkit for sustainable decision making in ITS deployment*. 2011. URL: http://www.2decide.eu.

[Bue04] BUERO WIDMER: *Verkehrsumlegungs-Modelle für stark belastete Strassennetze*. Frauenfeld, Switzerland, 2004. research assignment SVI 2001/541.

[Ede12] N. EDEN, A. TSAKARESTOS, I. KAPARIAS, A. GAL-TZUR, P. SCHMITZ, S. HAUPTMANN, and S. HOADLEY: "Using Key Performance Indicators for traffic management and Intelligent Transport Systems as a prediction tool". In: *19th ITS World Congress*. Vienna, 2012.

[Fro12] P. FROEHLICH and M. VRTIC: *Etablierung der Mikrosimulation des Verkehrs in der Stadt Zürich: Aufbereitung und Umlegung Makronetz*. Waedenswil, Olten, Switzerland: transSOL and transOPTIMA, 2012.

[Ge13a] Q. GE, B. CIUFFO AND M. MENENDEZ: "An exploratory study of two efficient approaches for the sensitivity analysis of computationally expensive traffic simulation models". In: *IEEE Transactions on Intelligent Transportation Systems* (2013) – under review.

[Ge13b] Q. GE and M. MENENDEZ: "An improved approach for the sensitivity analysis of computationally expensive microscopic traffic models: a case study of the Zurich network in VISSIM". In: *Proceedings of the 92th Annual Meeting of the Transportation Research Board*. Washington D.C., USA, Jan. 13–17, 2013.

[Ge13c] Q. GE and M. MENENDEZ: "An Efficient Sensitivity Analysis Approach for Computationally Expensive Microscopic Traffic Simulation Models". In: *International Journal of Transportation* (2013) – under review.

[Hei12] C. HEIMGARTNER: "Virtual ITS-Evaluation – Establishment of Micro-Simulation for the City Centre of Zurich". In: *19th ITS World Congress*. Vienna, 2012.

[Hua12] W. HUANG AND M. FELLENDORF: "Social Force Model for Vehicle Simulation on Operational Level". In: *19th ITS World Congress*. Vienna, 2012.

[ICT13] ICT-EMISSIONS CONSORTIUM: *Assessing the impact of ICT on road transport emissions*. 2013. URL: http://www.ict-emissions.eu.

[Ins13] INSTITUTE FOR LAND AND SEA TRANSPORT SYSTEMS, INSTITUTE FOR TRANSPORT PLANNING AND SYSTEMS AND SENOZON: *MATSim - Multi-Agent Transport Simulation*. 2013. URL: http://www.matsim.org.

[Kol03] T. G. KOLDA, R. M. LEWIS, and V. TORCZON: "Optimization by direct search: new perspectives on some classical and modern methods". In: *SIAM Review* 45.3 (2003), pp. 385–482.

[Mac67] J. MacQueen: "Some methods for classification and analysis of multivariate observations". In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 1967, pp. 281–297.

[Men12] M. Menendez and Q. Ge: "Final report on the calibration study for VISSIM (CSV)". Technical report. Zürich: ETH Zurich, 2012.

[Mul11] MULTITUDE: "State-of-the-art: Review of traffic data collection and estimation techniques and review of methodologies for traffic estimation, calibration and validation". MULTITUDE report, 2011.

[Ort12] J. Ortigosa, M. Menendez, and H. Tapia: "Study on the location of measurement points for an MFD perimeter control scheme in Zurich". In: *EURO Journal on Transportation and Logistics* (2012) – under review.

[Sal08] A. Saltelli, M. Ratto, T. Anres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola: *Global Sensitivity Analysis - The Primer*. Chichester, England: John Wiley & Sons, 2008.

[Sch13] L. Schiffmann, J. Ortigosa, and M. Menendez: "Measuring the MFD of Zurich – Identifying and evaluating strategies for efficient fixed monitoring resources". (2013) – working paper.

[Tec12] Technion Israel Institute of Technology, Imperial College London, Technical University of Munich, and Institute of Studies for the Integration of Systems and Polis: *CONDUITS DST, decision support tool*. Haifa, London, Munich, Rome, Brussels, 2012.

[Tom11] TomTom: *TomTom Traffic Stats, statistical traffic data*. Amsterdam: TomTom, 2011.

[Zur10] Canton of Zurich: *Transport model Canton of Zurich*. Zurich: Canton of Zurich, 2010.

[Zur11] City of Zurich: *On the way to the 2000-watt society, Zurich's path to sustainable energy use*. Zurich: Office for Environmental and Health Protection Zurich, City of Zurich, 2011. URL: http://www.stadt-zuerich.ch/content/dam/stzh/gud/Deutsch/Ueber%20das%20Departement/2000-Watt/Publikationen_und_Broschueren/OnTheWayToThe2000WattSociety.pdf.

Corresponding author: Christian Heimgartner, Modelling and Simulation, Division of Transport, City of Zurich, Switzerland. Phone: +41 44 411 88 60, e-mail: christian.heimgartner@zuerich.ch

Inflow-Regulating Traffic Light Control to Avoid Queue-Spillovers in Urban Road Networks

Stefan Lämmer, Martin Treiber, Markus Rausch
Technische Universität Dresden

Abstract

The capacity of a road network is predetermined by the green times at its intersections. As these are typically designed for expected traffic demands, spontaneous demand peaks or lane blockings will likely lead to congestion spreading over larger parts of the network, possibly ending up in a gridlock situation. We show that critical queue spillbacks can be avoided by an inflow-regulating traffic light control that does not serve more vehicles than subsequent roads can accommodate. In this way, vehicular queues only build up within designated areas of the roads segments, whereas upstream intersections remain fully accessible for non-affected flow directions. The accelerated propagation of segmented queues allows drivers to notice obstructions along their chosen route early enough to consider alternative routes or alternative modes of transport. In consequence, congested parts of the network are relieved from some traffic, whereas remaining capacities on surrounding roads are utilized. In a simulation study, the inflow-regulation principle noticeably prevented the emergence of a gridlock situation in two incident scenarios.

Keywords: vehicular traffic - road network - traffic light control - incident management - gridlock

1 Introduction

The signal timing of traffic lights is typically designed to provide the intersections of a road network with a throughput significantly higher than the average expected demand. A comprehensive overview of traffic light control strategies is given, for example, in Refs. [Por97; Hou01; Pap03]. Green times classically follow a cyclic scheme, with which a coordination of arterial roads is intended [Gar75]. In order to cope with variable traffic flows, some strategies, such as SCOOT [Bre04], dynamically adapt the control parameters. SCATS [Sim79], for example, recombines the pairs of intersections in coordination depending on actual demands.

More recent developments apply more advanced optimization techniques such as rolling-horizon methods [Por96], genetic algorithms [Bra08], and model predictive control [Lin12]. Nevertheless, the optimal control of traffic networks remains an unsolved problem of high computational complexity [Pap99]. Current research activities approach this complexity by designing rules that let desired dynamical properties such as synchronization or coordination emerge in a self-organized way. Relevant candidates are the "organic traffic light control" [Pro09], the "self-organizing traffic lights" [Ger12], the "schedule-driven control" [Xie12], or the "self-control" [Lam08]. A review on multi-agent approaches to traffic signal control is provided in [Che10].

The adaptation of intersection capacities towards variable traffic demands becomes problematic, however, in cases where the demand exceeds the maximum available intersection capacities or where queues start to spill back from one intersection to another. As soon as the growth of queues from cycle to cycle is inevitable, it becomes crucial to utilize remaining capacities in the most efficient way. An option would be to allocate maximum green times to those roads with the highest number of lanes at the cost of the other roads [Gaz02]. More generally, it is important to maximize the potential outflow from a critical region while restricting its inflow [Dag07]. Daganzo further distinguishes between "jam" and "gridlock" states. A gridlock state characterizes the precarious situation where an accumulation of vehicles in the network implies a blockade of potential outflow capacity, i.e. where the vehicles hinder themselves from leaving the network. Such gridlocks occur when vehicle queues spill back from one intersection to the next and eventually obstruct other flows. This might, in consequence, trigger a cascade. Therefore, efficient traffic light control requires an efficient prevention of queue spillovers.

This paper transfers Daganzo's principle of restricting the inflow into congested regions to a local traffic light control strategy, which reduces the amount of green times for those traffic flows that lead into congested road segments. Sec. 2 presents the basic idea and postulates particular features. Sec. 3 develops an analytical formulation of how traffic lights can avoid queue spillbacks by regulating its green times. Sec. 4 depicts the simulation study and its results. The key findings are concluded in Sec. 5.

2 Inflow-Regulation Principle

Inflow-regulation keeps the length of vehicle queues within certain bounds. As this local principle prevents intersections from gridlocks, it has several interesting implications on traffic flow at network scale. Some distinct features are illustrated in Fig. 1 and explained in the following.

2.1 Delimiting the Queue Length on a Road Segment

On a single road segment, congestion grows as soon as its inflow demand exceeds its outflow capacity. Extending the outflow capacity is often not possible, in particular if the downstream

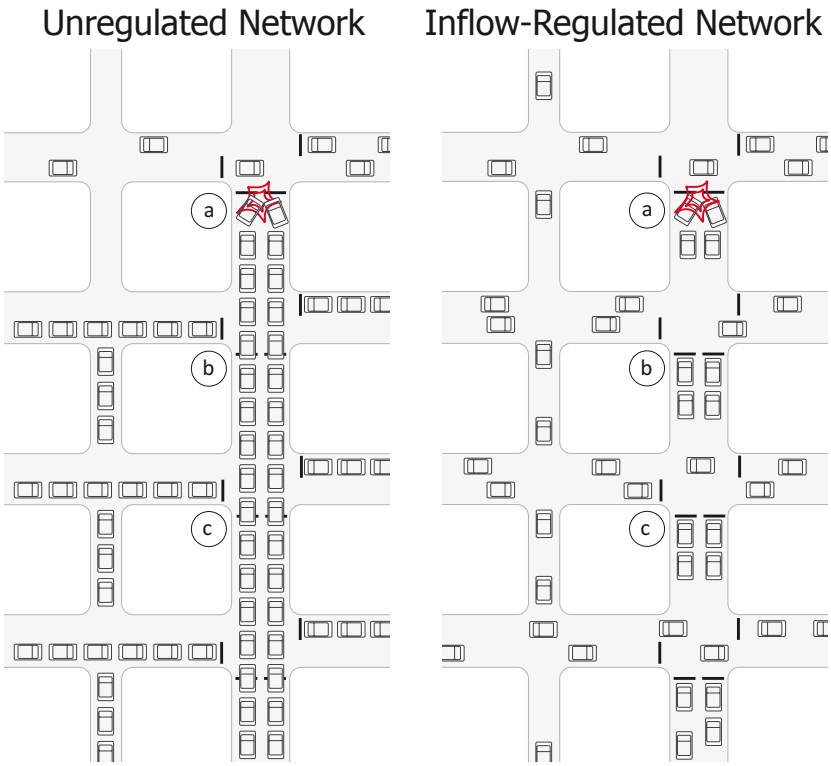


Figure 1: In a simple road network, where the main road is blocked due to an accident, the traffic situation evolves differently dependent on whether the intersections are inflow-regulated or not. Left: In the unregulated case, queues spill back and let larger parts of the network collapse. Right: (a) The purely local inflow-regulation hinders traffic from entering road segment before its queue exceeds a maximum length. Fewer cars encounter a standstill and the accident area remains accessible for rescue vehicles. (b) Intersections are not blocked and remain passable for unaffected flows. (c) The accelerated propagation of segmented queues lets drivers perceive the obstruction far ahead from the accident allowing them to decide for alternative routes.

intersection is over-saturated or if a lane is blocked. Instead, the upstream intersection can instantaneously limit the inflow into the congested road segment as soon as its queue tends to exceed a certain length. This can be accomplished by skipping or shortening associated green times. This imposes some requirements on the traffic light that controls the inflow into the regarded road segment. It must in particular be able to estimate the remaining queuing capacity of the road segment and how it evolves dependent on measured inflow and outflow rates. The analysis in Sec. 3 develops a formula with which conventional traffic light controls could be extended.

2.2 Keeping Intersections Passable for Unaffected Flows

As inflow-regulating traffic lights avoid the spillback of vehicle queues, the intersection remains passable for all flow directions that lead not into congested road segments. The in-

tersection remains accessible for other modes of transport as well as for emergency vehicles. Notice also, that drivers that intended to enter the congested road segment now face long red times. Since available turning directions are signalized with regular or even extended green times, however, drivers are able to consider alternative routes and use them.

2.3 Segmented Queue Growth Inhibits Global Gridlock Effects

Applying the inflow-regulation principle to multiple intersections in a network has several implications on how traffic and congestion evolves as a result of an incident. Most importantly, vehicular queues do not grow in a continuous manner. Instead, as they are restricted to build up within road segments only, queues grow segmentally. Since they spare out some space, segmented queues propagate faster towards the origins of traffic. While this means that drivers are held back earlier than in unregulated networks on the one hand, they are also able to notice obstructions further ahead on the other hand. Consequently, if some drivers decide to choose alternative routes or alternative modes of transport, traffic is redistributed such that remaining road capacities are utilized. Each vehicle that leaves a congested part of the network allows another vehicle to enter it a certain time later. One might observe that vehicle gaps start to propagate backwards from the exits to the entrances as analytically explained in [Hel06]. Congested parts of the network are relieved from traffic while unaffected flows can leave the network unrestrictedly. Since the emergence of gridlock situations is largely inhibited in regulated networks, the outflow capacity is only limited by the incident itself.

3 Traffic Light Control

The intersections of a road network are operated with traffic light controllers that, for example, optimize traffic flow with respect to minimum delays or vehicle stops. In order to incorporate the proposed inflow-regulation, the control algorithms are to extend by an additional rule saying *"Do not let more than $d_s(h)$ vehicles depart from traffic stream s within the next h seconds."* The following analysis develops a formula for $d_s(h)$.

3.1 Queue Model

Consider a single-lane road segment i as depicted in Fig. 2. It has a length of L_i meters, on which only the most downstream R_i meters are allowed for queues to build up. If vehicles require an effective space of l meters each in a queue, the number of queued vehicles is limited to R_i/l. In free traffic, vehicles travel at the maximum allowed velocity of v meters per second with which they would need L_i/v seconds to pass through.

The cumulated count of vehicles arriving and departing from road segment i at time t is denoted by $A_i(t)$ and $D_i(t)$, respectively. These numbers can be constructed, for example, from pulse counts of induction loops that are positioned at the upper and lower end of the road segment. Note, that there are two calibration conditions, $A_i(t - L_i/v) - D_i(t) = 0$ in case of free traffic, and $A_i(t) - D_i(t) = 0$ in case of an empty road, that allow to correct

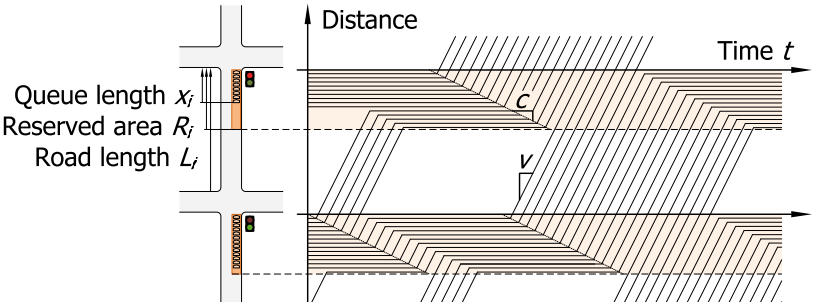


Figure 2: Left: The traffic lights along a road regulate the inflow into subsequent road segments such that vehicles only stop in reserved areas (shaded). Right: Associated vehicle trajectories indicate that maximum queue lengths are never exceeded. Vehicles are being served with a green light in a way, that they join subsequent queues at exactly those time points, at which they can fill in gaps departing vehicles created before.

detection errors. Also note, that, if the in- and outflow of a homogeneous road section is given, the temporal evolution of the queue length $x_i(t)$ can be analytically estimated with section-based methods, which assume flow continuity and a piecewise linear flow-density-relation [Dag94; Hel03; Tre13]. The following analysis is based on an estimate of the current queue length $x_i(t)$ and then balances the number of vehicles joining the queue against the number of vehicle-gaps that become available at the end of the queue after vehicles departed.

3.2 Anticipation of Queuing Capacity

At time t a queue of length $x_i \leq R_i$ has built up. The remaining $R_i - x_i$ meters of queuing space has a capacity for $(R_i - x_i)/l$ additional vehicles. Since this space is being partially filled with vehicles that arrived within the past $(L_i - x_i)/v$ seconds, there remains space for

$$\frac{R_i - x_i}{l} - \left[A_i(t) - A_i\left(t - \frac{L_i - x_i}{v}\right)\right] \tag{1}$$

additional vehicles to arrive.

Departures from the queue create gaps that allow subsequent vehicles to move up. Gaps propagate through the queue in opposite driving direction at a characteristic negative velocity $c \approx -5$ m/s. This means, the gap a vehicle created due to its departure at time $t + x_i/c$ compensates for another vehicle joining the queue at time t. The sum of gaps currently present within the queue, i.e. those that departing vehicles created within the past $-x_i/c$ seconds, can be filled with

$$D_i(t) - D_i\left(t + \frac{x_i}{c}\right) \tag{2}$$

vehicles reaching position x_i up to time t. If a vehicle arrives at road segment i at some future time point $t+h$, it will find a gap in the queue to fill in, if another vehicle has departed before time point $t + h + (L_i - x_i)/v + x_i/c$. Moreover, the sum of all vehicles that departed before

that time point will allow another

$$D_i\left(t+h+\frac{L_i-x_i}{v}+\frac{x_i}{c}\right)-D_i(t) \tag{3}$$

vehicles to enter the road segment up to time point $t+h$.

In order to answer the question, how many more vehicles

$$a_i(h) := A_i(t+h)-A_i(t) \tag{4}$$

are allowed to arrive at road segment i within time horizon h, the sum of Eqs. (1) to (3) has to be computed. As a result, the queue will not exceed a maximum length of R_i meters, as long as there will not arrive more than

$$a_i(h)=\frac{R_i-x_i}{l}+A_i\left(t-\frac{L_i-x_i}{v}\right)-A_i(t)+D_i\left(t+h+\frac{L_i-x_i}{v}+\frac{x_i}{c}\right)-D_i\left(t+\frac{x_i}{c}\right) \tag{5}$$

vehicles at road section i within the next h seconds.

3.3 Green Time Adjustment

At the intersection upstream of the investigated road segment i, there are several incoming traffic streams s. Ideally, each stream would lead to exactly one outgoing road segment i, which is the case for separate turning lanes. In more general cases with mixed lanes, however, there will only a certain fraction α_{si} of the vehicles of traffic stream s turn into road segment i. Obviously, $\sum_i \alpha_{si}=1$ is valid for all streams s. Since the number of vehicles that arrive at i is restricted by Eq. (5), the maximum number of vehicles allowed to depart from stream s within horizon h follows to be:

$$d_s(h)=\min_i \frac{a_i(h)}{\alpha_{si}} \tag{6}$$

In order to ensure that the queues on neither of its outgoing roads i exceed a certain critical length R_i, the corresponding traffic light control must not allocate more green time to its incoming traffic streams s within time horizon h than $d_s(h)$ vehicles require to depart.

3.4 Discussion

In the more general case of multi-lane traffic, the effective length l of a vehicle in a queue has to be divided by the number of lanes. Note also that the turning fractions α_{si} will vary in time as soon as drivers consider alternative turning directions. Even if Eq. (6) only serves as a rough estimate, using it with historical turning fractions obtained from non-perturbed traffic situations is still safe with respect to a reliable inflow-regulation. As drivers tend to avoid congestion, historical α_{si} values overestimate the actual turning rate into congested road segments i, which results in rather under-critical green times for the corresponding streams

s. Another practical issue is that safety directives may impose maximum red times. In these cases, at least a short green light must be given after a certain while. Furthermore, the time shift $h + (L_i - x_i)/v + x_i/c$ in Eq. (5) might become positive, which implies a reference to future departures at the subsequent intersection. If the signal timing of that intersection is not known in advance, a safe measure could be to assume that future departures will not take place, i.e. to assume $D_i(t') = D_i(t)$ for future time points $t' > t$. The available queuing capacity will thereby be underestimated. In a scenario, however, with $L_i = 500$ m, $v = 15$ m/s, $c = -5$ m/s, and $x_i = 200$ m, the respective time shift will not become positive for horizons $h \leq 20$ s.

A particularly well suited control strategy to be combined with the proposed inflow-regulation is the "Self-Control" as introduced in [Lam08]. It is not restricted to cycle times or to a fixed order of phases. It does instead calculate a priority index for each traffic stream based on the short-term anticipation of the number of cars expected to arrive in the queue within the next few seconds, and gives green to those non-conflicting streams for which the priority index is highest. In case of variable inflows, this naturally results in irregular switching sequences, in which some streams might be served more often than others. An implementation of the inflow-regulation principle would require to calculate the priority indices over no more than $d_s(h)$ vehicles according to Eq. (6). If a particular stream leads into a queued road, the corresponding priority index will drop, and green times are automatically allocated to other streams.

4 Simulation Results

The proposed inflow-regulation principle was validated by using the example network depicted in Fig. 1. The simulation runs were accomplished with the commercial traffic flow simulation tool PTV Vissim (Version 5.40). The road segments between two intersections have a length of 70 m. Roads on which traffic enters the network, however, were chosen long enough to accommodate all queues that result from the regarded incidents. The inflow traffic volume was set to 1200 veh/h on the main road and to 300 veh/h on the others. Turnings were generally not permitted. The intersections were operated with fixed-time controllers and a common cycle time of 60 s. Green times were set to 40 s for the main road direction and to 20 s for the other flow directions. The offsets were chosen to let a green wave propagate along the main road. The inflow-regulating principle was implemented such that a scheduled green time is given only if the vehicle queue on downstream road segments does not exceed a maximum length of 35 m. Saved green times were not redistributed to other flow directions.

Two incident scenarios, A and B, are considered. Both reflect an accident near the most downstream stop line of the main road. In scenario A, drivers can bypass the accident via one lane at walking speed. In scenario B, the accident blocks both lanes of the road. Both incidents are active for a duration of 800 s.

For a quantitative comparison of the regulated and unregulated network, Fig. 3 shows how the accumulation of vehicles N evolves in each case. As the regarded incidents cause a

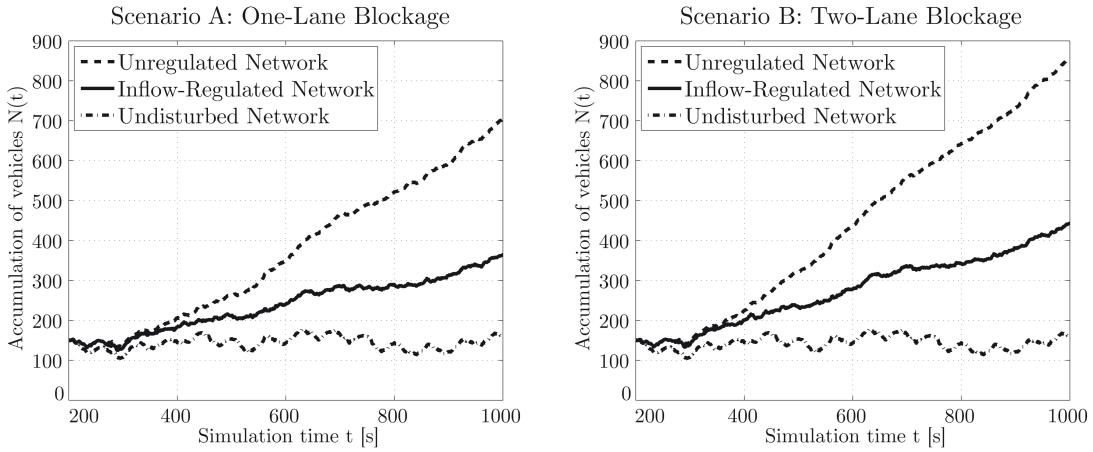


Figure 3: For both scenarios A (left) and B (right), the accumulation of vehicles is significantly lower in the inflow-regulated network (solid) as compared to the unregulated network (dashed). This is due to the fact that inflow-regulated intersections avoid the spillover of queues and, thereby, remain fully passable for all non-affected flows. In both scenarios, the outflow capacity of the regulated network is only limited by the incident itself.

queue to grow along the main road in any case, the difference is on how much non-affected flows were involved. The unregulated network resulted in a distinct gridlock situation, in which eventually no vehicle was able to leave the network anymore. Contrarily, the inflow-regulation principle prevented the main road queues from spilling back to upstream intersections such that the flows from the other roads were not obstructed by the incident. In consequence, the number of vehicles in the network could be reduced by 48.4 %. Results are summarized in Table 1.

Table 1: Accumulation of vehicles N at the end of each simulation run ($t = 1000\,\mathrm{s}$).

	Scenario A	Scenario B
Unregulated Network	703	857
Inflow-Regulated Network	363	442
Relative savings	48.4 %	48.4 %

5 Conclusion

The proposed inflow-regulation principle prevents vehicle queues from spilling back to the most upstream intersection. Traffic lights regulate the inflow into congested road segments by shorter green times to make sure its queue does not exceed a certain predefined length. This purely local principle has several implications. First of all, the corresponding intersections remain fully accessible for traffic with different destinations and, additionally, for traffic

that bypasses congestion. Hence, precarious gridlock effects are largely inhibited and the impact of the incidents is limited to a much lesser extent. Taking all effects together, the inflow-regulation principle manages incidents in a much faster and more efficient way.

Only being allowed to build up on road segments, queues grow segmentally. This implies that information of obstructions along a certain route propagates faster through the network. This gives drivers a chance to consider and choose alternative routes soon enough. However, drivers may experience that the traffic lights along a chosen route are red for an exceptionally long time. Neither the reason of this obstruction is obvious, nor will the drivers by themselves be able to estimate expected travel times along alternative routes, nor will they know what routes are accessible at all. Consequently, each individual driver will face a complex decision process in which he must heuristically decide to either stay on the original route for a certain while, or to move on to one of the potentially available turning directions. Online navigation systems [Coh09] might support this decision process.

The simulation study has resulted in a significant reduction of the accumulation of vehicles in the inflow-regulated network. Vehicles of other flow directions were not hindered from crossing the intersections and, thus, from leaving the network. The regulated network, therefore, was relieved from some traffic and could also avoid the emergence of a gridlock situation. In the simulation study, drivers were not allowed to turn. It remains a future task to develop an according route choice model of individual drivers. It is to expect that the redistribution of traffic among available bypass roads will further decrease the accumulated number of vehicles in the network.

We further propose to extend the inflow-regulation principle by additionally reallocating the green times of possibly present turning directions. While the green times for flows into congested roads are shortened or skipped, the green times for available turning directions could be likewise expanded. This provides capacity to those drivers that decide for an alternative route. Both, the inflow-regulation principle and the instantaneous reallocation of green times, give rise to a purely autonomous capacity regulating control concept in partially obstructed networks. If it was possible to redistribute all affected traffic flows along the remaining network capacities, the network would have "healed" itself. Several, more theoretical, questions remain to be addressed in further studies. These questions include how non-equilibrium traffic states in networks with capacity regulating traffic lights can be modeled, and what preconditions a network must have in terms of road capacity, traffic load, or routing alternatives, in order to satisfy given traffic demands also in case of incidents.

Acknowledgment

The authors gratefully thank the German Research Foundation (DFG-project Tr 1102/1-1) for supporting this research. Thanks also to Christian Richter for accomplishing the simulation runs.

References

[Bra08] R. BRAUN, C. KEMPER, and F. WEICHENMEIER: "TRAVOLUTION – Adaptive urban traffic signal control with an evolutionary algorithm". In: *Proceedings of the 4th International Symposium "Networks for Mobility"*. Stuttgart, 2008.

[Bre04] D. BRETHERTON, M. BODGER, and N. BABER: "SCOOT - the future". In: *12th IEE International Conference on Road Transport Information and Control*. Vol. 501. 2004, pp. 301–306.

[Che10] B. CHEN: "A Review of the Applications of Agent Technology in Traffic and Transportation Systems". In: *IEEE Transactions on Intelligent Transportation Systems* 11.2 (2010), pp. 485–497.

[Coh09] N. COHN: "Real-Time Traffic Information and Navigation - An Operational System". In: *Transportation Research Record: Journal of the Transportation Research Board* 2129 (2009), pp. 129–135.

[Dag07] C. F. DAGANZO: "Urban gridlock: Macroscopic modeling and mitigation approaches". In: *Transportation Research Part B* 41.1 (2007), pp. 49–62.

[Dag94] C. F. DAGANZO: "The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory". In: *Transportation research B* 28.4 (1994), pp. 269–287.

[Gar75] N. GARTNER, J. LITTLE, and H. GABBAY: "Optimization of Traffic Signal Settings by Mixed-Integer Linear Programming Part I: The Network Coordination Problem". In: *Transportation Science* 9.4 (1975), pp. 321–343.

[Gaz02] D. C. GAZIS: *Traffic Theory*. International Series in Operations Research & Management Science. Springer, 2002.

[Ger12] C. GERSHENSON and D. A. ROSENBLUETH: "Self-organizing traffic lights at multiple-street intersections". In: *Complexity* 17.4 (2012), pp. 23–39.

[Hel03] D. HELBING: "A section-based queueing-theoretical traffic model for congestion and travel time analysis in networks". In: *Journal of Physics A* 36 (2003), pp. L593–L598.

[Hel06] D. HELBING, A. JOHANSSON, J. MATHIESEN, M. H. JENSEN, and A. HANSEN: "Analytical approach to continuous and intermittent bottleneck flows". In: *Physical Review Letters* 97 (2006).

[Hou01] N. B. HOUNSELL and M. MCDONALD: "Urban network traffic control". In: *Journal of Systems & Control Engineering* 215.4 (2001), pp. 325–334.

[Lam08] S. LÄMMER and D. HELBING: "Self-Control of Traffic Lights and Vehicle Flows in Urban Road Networks". In: *Journal of Statistical Physics* (2008), P04019.

[Lin12] S. LIN, B. DE SCHUTTER, Y. XI, and H. HELLENDOORN: "Efficient network-wide model-based predictive control for urban traffic networks". In: *Transportation Research Part C: Emerging Technologies* 24 (2012), pp. 122–140.

[Pap03] M. PAPAGEORGIOU, C. DIAKAKI, V. DINOPOULOU, A. KOTSIALOS, and Y. WANG: "Review of road traffic control strategies". In: *Proceedings of the IEEE* 91.12 (2003), pp. 2043–2067.

[Pap99] C. H. PAPADIMITRIOU and J. N. TSITSIKLIS: "The Complexity of Optimal Queuing Network Control". In: *Mathematics of Operations Research* 24.2 (1999), pp. 293–305.

[Por96] I. PORCHE, M. SAMPATH, R. SENGUPTA, Y.-L. CHEN, and S. LAFORTUNE: "A decentralized scheme for real-time optimization of traffic signals". In: *Proceedings of the 1996 IEEE International Conference on Control Applications*. 1996, pp. 582–589.

[Por97] I. PORCHE and S. LAFORTUNE: "Dynamic traffic control: decentralized and coordinated methods". In: *IEEE Conference on Intelligent Transportation System*. 1997.

[Pro09] H. PROTHMANN, J. BRANKE, H. SCHMECK, S. TOMFORDE, F. ROCHNER, J. HAHNER, and C. MÜLLER-SCHLOER: "Organic traffic light control for urban road networks". In: *International Journal of Autonomous and Adaptive Communications Systems* 2.3 (2009), pp. 203–225.

[Sim79] A. G. SIMS: "SCATS: The sydney co-ordinated adaptive system". In: *Proceeding of the Engineering Foundation Conference on Research Priorities in Computer Control of Urban Traffic Systems*. 1979.

[Tre13] M. TREIBER and A. KESTING: *Traffic Flow Dynamics*. Data, Models and Simulation. Springer, 2013.

[Xie12] X.-F. XIE, S. SMITH, and G. BARLOW: "Schedule-Driven Coordination for Real-Time Traffic Network Control". In: *Proceedings 22nd International Conference on Automated Planning and Scheduling*. 2012.

Corresponding author: Stefan Lämmer, Technische Universität Dresden, "Friedrich List" Faculty of Transportation and Traffic Sciences, 01062 Dresden, Germany, phone: +49 351 463 36802, e-mail: stefan.laemmer@tu-dresden.de

Integrating Weather Impact in Travel Demand Models for Private Motorised Transport

Bernhard Heilmann[1], Martin Reinthaler[1], Johannes Asamer[1], Lena Fehrenbach[2], Juliane Pillat[3] , Jochen Lohmiller[3], Markus Friedrich[3], Karl Schedler[4]

[1] Austrian Institute of Technology

[2] UBIMET

[3] Universität Stuttgart

[4] micKS

Abstract

Travel demand models describing demand-supply-interaction of individual transport are important building blocks of transport planning and traffic management systems. Weather conditions, e.g. top weather or adverse weather like rainfall or snow, can affect several levels of demand-supply-interaction. In order to quantify weather impact on travel demand models, traffic and weather data from an inter-urban motorway in Bavaria, Germany and the intra-urban road network in the Austrian capital Vienna have been analysed in the framework of a cooperative research project.

By combining absolute and relative thresholds, which were defined relative to local climate summary values, standardised weather classes were defined.

On the inter-urban motorway, weather impact on travel demand differed among weekdays. Demand increase for days with top weather and demand decrease for rainy and snowy days was quantified. In the intra-urban road network, weather impact depended mainly on the type of precipitation. Whereas snow decreased passenger car demand, rain and even heavy rainfall had no significant impact on demand.

On the inter-urban motorway, free speed reduction and capacity reduction due to heavy rain and snowfall was quantified. In the intra-urban road network, the network flow-speed relation was similar for dry road, wet road combined with rain and slush combined with snow. Only for snowy road combined with snow, lower network speeds were observed for given flows.

Based on the quantified weather impacts, recommendations are given on how to integrate weather impact in travel demand models.

Keywords: travel demand model, weather impact

1 Objectives

Travel demand models describing demand-supply-interaction of private transport are important building blocks of transport planning and traffic management systems. Commonly travel demand is represented by a table of (estimated) origin-destination flows. An assignment model is used to assign the OD flows to routes, which are propagated on links along the routes. A performance function is used to calculate flow-dependent supply costs, which are frequently measured as travel time delays. Weather conditions, e.g. top weather or adverse weather like rainfall or snow can affect several levels of demand-supply-interaction. The severity of the impact depends on other factors (e.g. day of year, time of day, traffic composition).

Whereas a variety of methods have been described for calibration of travel demand models in general (see i.a. [Cas09]), an integrated approach for model calibration taking into account weather impact on all levels of demand-supply-interaction could not be directly derived from existing studies and methods (see e.g. [Che12], [Liu13], [Rak07], [Zha09]).

A comprehensive summary of weather impact on macroscopic and microscopic traffic characteristics, on traffic safety as well as on road surface conditions has been elaborated by COST Action TU0702 “Real-time monitoring, surveillance and control of road network under adverse weather conditions” (see [Fau11], pp. 21 – 47).

The evaluated studies targeting impact of weather on traffic demand have reported a wide impact range. Impact of rain on demand ranged from no measurable effect to about 4 % decrease on working days and from 20 % increase to 20 % decrease on weekends. Differences could be due to the population’s different reaction to local climate for each trip purpose or differing availability of transport mode (especially availability of public transport facilities in urban areas). Whereas the impact of adverse weather on driving speed is comparable for different regions, analysis revealed large ranges for road capacity (e.g. between about 5 and 20 % for light rain). This can be attributed to missing comparability of the actual road surface state. In order to improve comparability of results, standardisation of weather classifications for demand analysis taking local climate into account as well as for speed and performance analysis have been identified as an important objective.

Classifications of road weather situations and methods for demand and supply performance analysis are introduced in chapter 2. As a basis for quantifying weather impact on travel demand models in a standardised way, traffic and weather data from the inter-urban motorways A8 and A93 in Bavaria, Germany and the intra-urban road network in the Austrian capital Vienna have been analysed in the framework of a cooperative research project. The main results of this project are presented in chapter 3. Based on the quantified impacts, recommendations are given in chapter 4 on how to integrate weather impact in travel demand models.

2 Methods

2.1 Classification of road weather situations

As a prerequisite for impact analysis, a separate weather classification for demand analysis and for speed and performance analysis were elaborated.

For demand analysis, daily weather values of the region under investigation were used. As the Bavarian region and the Viennese region belong to different climates, resulting in different yearly amount of precipitation, a combination of absolute and relative thresholds, which were defined relative to local climate summary values, were applied for classification (see Table 1).

Table 1: Classification for impact analysis on demand

	sunshine	**precipitation type**	**precipitation intensity (mm/d water equivalent)**
dry	-	-	-
top weather	>= 90% of max. sunshine duration	-	<= 0.2 mm/d
snow	-	snow	>= 1mm/d
rain	-	rain	> 0.2 mm/d
heavy rain	-	rain	> 95th perc. (of monthly cumulative distribution)

A threshold for heavy rain was defined as the 95th percentile of the monthly cumulative distribution of daily rainfall. This threshold for heavy rain therefore differed between regions and between seasons.

For the analysis of weather impact on traffic speed and performance, a separate weather classification was applied (see Table 2 and [Kir12]). As the immediate impact of the local road surface condition on driving behaviour was investigated, classification thresholds were defined for 15 min aggregated road weather sensor data.

2.2 Methods for demand analysis

Demand on the inter-urban motorway (Federal Motorways A8 and A93)

Travel demand in the Bavarian survey area is influenced by a couple of factors. On the one hand the motorways A8 and A93 are typically used for intra-European holiday trips in summer and in winter, with peaks during the country-specific school holidays, but also outside of these periods (seasonal effect and holiday effect). On the other hand many short-trips on Fridays and long weekends undertaken by regional travelers could be observed (special days). To isolate these effects from weather impact, a three-fold analysis was done:

- A linear regression analysis was applied for every count location and weekday separately. The regression function integrated the main factors, especially weather, influencing daily traffic volume.

Table 2: Classification for impact analysis on supply performance.

	precipitation type	precipitation intensity (mm/h water equivalent)	wet bulb temperature (°C)	road surface temperature (°C)
dry	-	= 0	*	*
wet road / rain	rain	< 0,5	>= 0	> - 2
slippery road / rain	rain	< 0,5	>= 0	<= - 2
wet road / strong rain	rain	>= 0,5	>= 0	> - 2
slush / snow	snow	< 0,5	< 0	> - 2
snowy road / snow	snow	< 0,5	< 0	<= - 2
slush / medium snow	snow	>= 0,5; < 3,5	< 0	> - 2
snowy road / medium snow	snow	>= 0,5; < 3,5	< 0	<= - 2
slush / heavy snow	Snow	>= 3,5	< 0	> - 2
snowy road / heavy snow	Snow	>= 3,5	< 0	<= - 2

- A cluster analyses was done, to examine weather impact on load curves of traffic flow.
- A household survey was evaluated to derive typical travel behaviour parameters like trip production rates and trip distance distribution according to several weather situations.

The ANPR (Automated Number Plate Recognition) systems load curves in the survey area provided additional information about the local origin of the detected vehicles. Hence regression and cluster analysis was applied to total traffic flows as well as to partial traffic flow with certain origins.

Demand on the intra-urban road network (city of Vienna)

In the city of Vienna, demand was measured at 58 measurement sites, which were located on urban arterial roads with a free speed level between 30 and 60 km/h. Demand flow was measured separately for passenger cars and heavy goods vehicles (including all larger vehicles) and summed up for all measurement sites during a one hour interval, resulting in an intra-urban demand flow level. Distribution of demand flow levels was compared among the weather classes, applying a Wilcoxon rank sum test. In addition, daily time series plots for different weather classes were compared, in order to check for temporal effects as e.g. during rush hours.

2.3 Methods for supply performance analysis

Supply performance on the inter-urban motorway (Federal Motorways A8 and A93)

In order to determine weather impact on speed and capacity at cross sections of motorways

A8 and A93, road weather data combined with traffic data were analysed in the period from 2008 to 2012. Road weather data (road surface temperature, road condition, type and intensity of precipitation, air temperature and humidity) were measured in 1 minute intervals. Traffic data (average speed, traffic flow and average time gap per 1 minute interval) were acquired separately for each lane and classified into passenger cars and trucks.

In order to determine vehicle free speed and capacity, data of selected cross sections were aggregated in 15 min intervals and classified according to the defined road weather classes. Two basic models were used to describe the fundamental diagram for each road weather class. For macroscopic modeling, the traffic flow – vehicle speed relation described by Van Aerde [VAe95] was used. The Van Aerde equation parameters for each weather class were computed by regression analysis. For microscopic modeling of weather impact, the fundamental diagram model described by N. Wu [Wu00] was applied. The parameters of the phase equations of the Wu model were computed by a regression method (ordinary least square estimator) for each road weather class.

In order to enhance the accuracy of the modelling, a continuous model was developed, which avoids using weather classes. The input of the model uses predictable weather parameters like air- and dew-point temperature, precipitation as well as road surface temperature. The model structure was derived from basic principles of driving dynamics, tire friction and visibility resp. stopping distance and certain road construction features. The parameters of the model were estimated by multiple regression methods. The model outputs free speed and capacity reduction corresponding to road weather data (see [Sch12]).

Supply on the intra-urban road network (city of Vienna)

On intra-urban road links, measurement of performance functions (e.g. volume-delay functions) proves to be difficult due to instationary traffic flow. Network flow-speed relations show a more stable behaviour than flow-speed relations on single links (see e.g. [Gero2010]). To this aim, each link of the Viennese road network was grouped according to its free speed (20 to 30 km/h up to 80 to 90 km/h). Within each group, harmonic average network speed for a one hour interval was calculated. Network speeds of urban arterial roads with a free speed level between 30 and 60 km/h and average demand flow were compared in a network flow-speed relation (“network fundamental diagram”).

3 Results

3.1 Demand on the inter-urban motorway (Federal Motorways A8 and A93)

On motorway A8 between Munich and Salzburg, weather impact differed locally and among weekdays. On a representative count location (see Figure 1), weather impact on daily traffic flow could be observed on Friday, Saturday and Sunday (see Table 3). The strongest impact could be observed on Sundays. Moreover, snow decreased traffic flow more than rain.

Table 3: Deviation of weekday specific average daily traffic flow [%].

	Top weather	Rain	Snow
Friday	+5%	-3%	-17%
Saturday	+4%	-8%	-21%
Sunday	+6%	-10%	N.A.

The left part of Figure 1 illustrates rain impact for different locations in the network. Analogously to top weather, rain impact decreases with increasing distance to the regional capital Munich. The right part shows the shift of the origin-specific daily traffic volume detected on the ANPR location. The detected traffic volume of vehicles is increasing for number plates registered in administrative districts located in the South-Eastern region of Munich, which indicates weather related origin-destination choice.

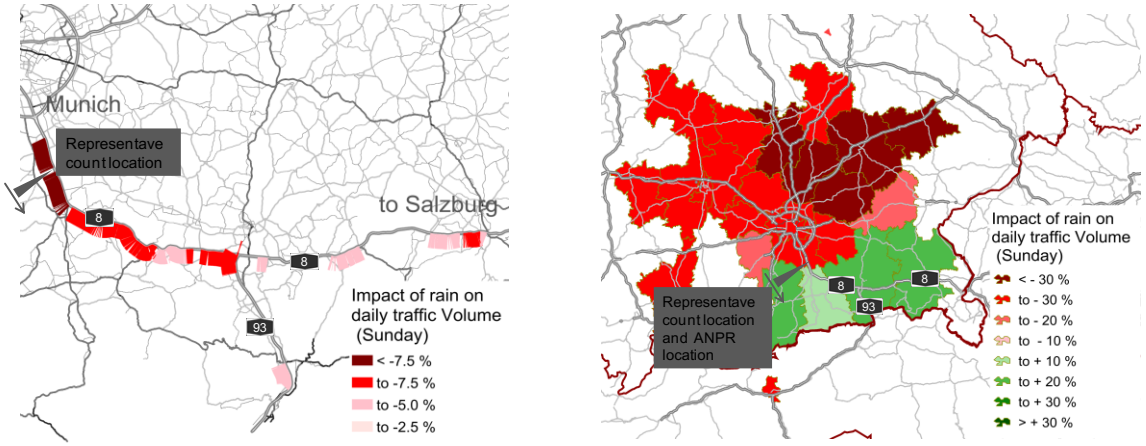


Figure 1: Percentage shift of daily traffic volume on Sunday due to rain.

The evaluation of the household survey confirms the weather related travel behaviour: Weather influences the number of trips and also the resulting trip distance distribution. The average trip length increases during top weather and decreases during rain. Furthermore cluster analysis reveals that the load curves of local traffic are also affected by current weather situations.

3.2 Demand on the intra-urban road network (city of Vienna)

Top weather increased passenger car demand by 6% on working days during vacation periods, whereas it did not increase passenger car demand during holidays. Top weather also increased heavy goods vehicle demand on working days during vacation periods. As the share of heavy goods vehicles constituted less than 5% of intra-urban traffic flow, the impact of heavy goods vehicles was considered to be small.

Rainfall and also heavy rainfall (exceeding the 95% percentile of the monthly distribution) had no significant impact, neither on passenger car demand nor on heavy goods vehicle demand. Snow resulted in an 11% reduction of passenger car demand on working days and a

Table 4: Comparison of weather impact on travel demand.

	Intra-urban road network (Vienna)	Inter-urban motorway (A8, A93)
Target of analysis	Hourly traffic flow level of 58 detectors	Daily traffic flow of single detectors
Preclassification	Day classes (working day, vacation period, holiday), vehicle class	Day classes (weekdays)
Method	Comparison of distribution with rank sum test	Regression analysis Cluster analysis Evaluation of household survey
Top weather impact	Increase on working days during vacation periods	Increase, largest impact on Saturdays and Sundays
Rain impact	No impact	Reduction, largest impact on Saturdays and Sundays
Snow impact	Reduction on working days and holidays	Reduction larger than during rain, largest impact on Saturdays
Additional remarks	similar impact in the whole city	Different impact in different locations

17% reduction of passenger car demand on holidays. In the same way, snow resulted in a decrease of heavy goods vehicle demand on holidays (-13%) and most likely on working days (-11%).

The main findings of demand analysis in both regions are summarised in Table 4.

3.3 Supply performance on the inter-urban motorway (Federal Motorways A8 and A93)

On the Federal Motorway A8 between Munich and Salzburg, archived road weather and traffic data from 48 measurement sites in the period from 2008 until 2012 were evaluated. An explicit clustering of the flow-speed relation as a function of the weather classes could be observed (see Figure 2). During heavy rain and increased waterfilm depth, the distribution of vehicle free speeds was reduced considerably, while effect on capacity was minor. During snowfall and snow-covered road surface, a significant decrease of chosen vehicle free speed and also of capacity was encountered.

For cross sections with two lanes in each direction and no significant longitudinal and transversal slope, the following average reduction factors for each observed road weather class could be calculated (see Table 5).

Based on measured route travel times from 2011, reliability of travel times and causes of delays on three sections of the inter-urban motorway were investigated. Increased demand was found to be the main cause for delays (83% of cases). In 3% of cases, rain could be identified as the only cause of delays, in 7% of cases a combined impact of travel demand and rain was detected.

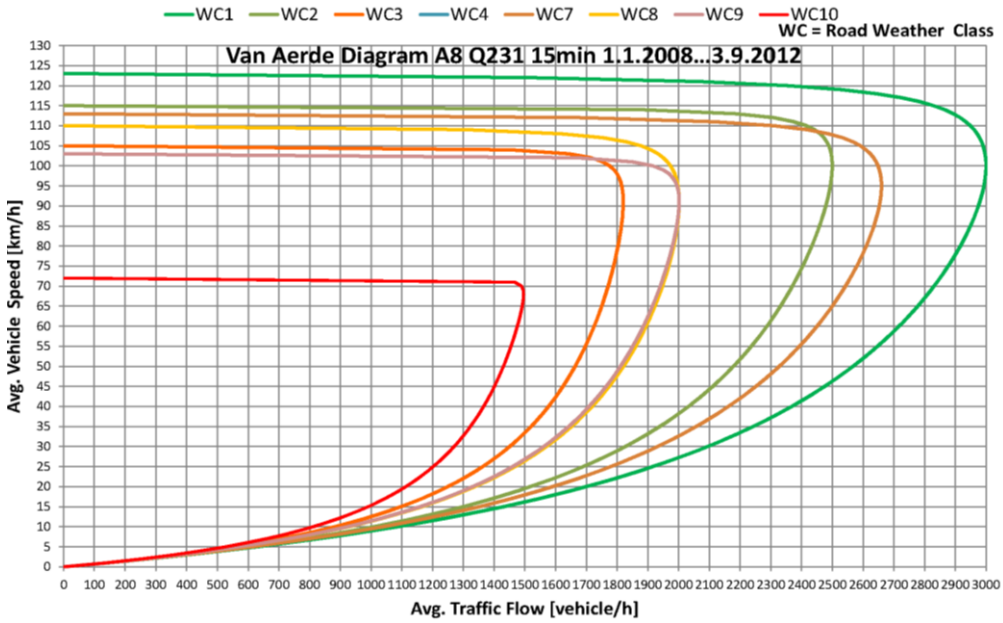


Figure 2: van Aerde flow – speed diagram for different weather classes.

3.4 Supply performance on the intra-urban road network (city of Vienna)

Network speeds of urban arterial roads with a free speed level between 30 and 60 km/h and average demand flow were combined in a network flow-speed relation. For each speed category, a linear decrease of network speed with increasing average demand flow could be observed. This relation was similar for dry road, wet road combined with rain and slush combined with snow. For snowy road combined with snow, the relation was shifted towards lower network speeds for given flows.

In spite of the clear shift of the flow-speed relation during snowy road combined with snow, an impact of this road weather situation on network speed could not be detected. Only the modes of network speed distributions were lowered. This small impact is likely due to a compensatory effect by reduced total passenger car flow on snowy days.

Table 5: Observed reduction factors due to road weather condition on 2-lane motorway.

	Class number	Free Speed Reduction	Capacity Reduction
dry	1	0%	0%
wet road / rain	2	-31%	0%
slippery road / rain	3	-54%	-18%
wet road /strong rain	4	-39%	-15%
slush / snow	5	-31%	-3%
snowy road / snow	6	-39%	-15%
slush / medium snow	7	-57%	-40%
snowy road / medium snow	8	-67%	-15%
slush / heavy snow	9	-67%	-54%
snowy road / heavy snow	10	-69%	-57%

For wet road combined with rain, network speeds were shifted towards lower speeds, resulting in a small reduction of median network speed of less than 3 %.

The main findings of supply performance analysis in both regions are summarised in Table 6.

Table 6: Comparison of weather impact on traffic supply performance.

	Intra-urban road network (Vienna)	Inter-urban motorway (A8, A93)
Target of analysis	Network speed, traffic flow - speed relation	Density – speed relation, traffic flow – speed relation, travel time reliability
Measurement values	58 stationary sensors, taxi travel times for entire road network	48 stationary sensors, 4 ANPR cameras
Method	Comparison of network speed per road class and average traffic flow	Local analysis of sensor data, travel time distribution from ANPR data
weather impact on speed	Small reduction of network speed on wet road during rain	Reduction of free speed during heavy rain (increased water film depth), snow and snowy road
Weather impact on performance /capacity	Impact of snowy road combined with snow on flow – speed relation	Capacity reduction during snowfall and snowy road minor capacity reduction during rain
Weather impact on travel time reliability	no investigation	Small reduction during rain
Model results	Traffic flow – travel time relations for different road classes and road weather classes	Density – speed relations, traffic flow – speed relations for different road weather classes factors for travel time reliability

4 Discussion and recommendations

For each day class and weather class with proven impact on supply-demand interaction, either the travel demand model or the traffic supply model or both have to be re-calibrated.

4.1 Re-calibration of demand model

Trip generation, trip distribution, time-choice or mode choice within the travel demand model could be the target of re-calibration. As further information is needed to be able to calibrate the right level of the demand model, the calibration depends on the available data sources. Table 7 shows the application of available data sources to integrate weather impact on several demand model levels.

The result of re-calibration is an OD matrix for different day classes in combination with different weather classes (e.g. for different weekdays combined with dry weather, top weather, rain and snowfall).

Table 7: Application of different data sources to calibrate and validate the demand model.

	Derivable parameters for different weather classes	Possible validation level
ANPR-Data	destination-specific load curves (directly)	trip generation trip distribution
Mobile phone data	parameters for destination choice model (directly) parameters of mode choice model (directly)	trip distribution mode choice (route choice)
Data from household survey	trip production rates (directly) parameters for destination choice model (indirectly) parameters of mode choice model (indirectly)	trip generation trip distribution mode choice
Traffic volumes of local detectors	-	route choice / assignment

4.2 Recalibration of supply model

Inside the supply model, the performance function is re-calibrated for different weather classes. Based on the estimated fundamental diagrams for each weather class, free speed reduction and capacity reduction is estimated for each weather class. The reduction factors for free speed and capacity are applied to the performance function in the supply model, resulting in different performance functions for different weather classes. On intra-urban road links, network flow-speed relations show a more stable behaviour than flow-speed relations on single links. Therefore, observed network flow-speed relations for different road classes can be used as substitutes for flow-speed relations on individual links.

After recalibrating demand model and supply model, prediction capability of the adapted model has to be validated. For this purpose, the model is used for prediction of traffic state or travel time during a defined validation period in the investigated road network. During this validation period, all weather situations with proven impact on supply-demand interaction should be included.

Acknowledgement

The research work was organized as trans-national project with participation of German and Austrian partners and co-funded by the Austrian Ministry of Transport, Innovation and Technology (BMVIT) and the German Federal Ministry of Economy and Technology (Bundesministerium für Wirtschaft und Technologie).

References

[Bac06] K. BACKHAUS, B. ERICHSON, W. PLINKE, and R. WEIBER: *Multivariate Analysemethoden*. 11. Ed. Heidelberg, Germany: Springer, 2006.

[Cas09] E. CASCETTA: *Transportation Systems Analysis*. Springer, 2009.

[Fau11] N. E. EL FAOUZI (ED): *Real-time monitoring, surveillance and control of road networks under adverse weather conditions. Effects of weather on traffic and pavement: State of the art and best practices*. 2011. COST Action TUO702. ISBN: 978-2-85782-688-0

[Che12] R. B. CHENG and K. J. CLIFTON: "Traveling in Comfort: Investigating Weather Ranges for Travel". In: *91st Annual Meeting of the Transportation Research Board*. Washington, D.C., 2012.

[Ger08] N. GEROLIMNIS and C. F. DAGANZO: "Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings". In: *Transportation Research Part B* 42 (2008), pp. 759–770.

[Kir12] H. KIRSCHFINK, M. POSCHMANN, D. ZOBEL, and K. E. SCHEDLER: „Stauprävention im Winterdienst" In: BUNDESANSTALT FÜR STRAßENWESEN (ED.): *Verkehrstechnik.* Vol. V 215 (2012).

[Liu13] C. LIU, Y. O. SUSILO, and A. KARLSTRÖM: "Investigating the impacts of weather variability on individual's daily activity-travel patterns: a comparison between commuters and non-commuters in Sweden" In: *93rd Annual Meeting of the Transportation Research Board*. Washington, D.C., Jan. 29, 2014. – paper submitted for presentation.

[Sch12] K. E. SCHEDLER: "New Approaches for Modelling of Weather impact on Traffic flow". In: *SIRWEC, 16th International Road Weather Conference*. Helsinki, Finnland, May 23–25, 2012.

[Rak07] H. RAKHA, M. FARZANEH, M. ARAFEH, R. HRANAC, E. STERZIN, and D. KRECHMER: *Empirical Studies on Traffic Flow in Inclement Weather*. Final Report - Phase 1. 2007.

[VAe95] M. VAN AERDE: "A single regime speed-flow-density relationship for freeways". In: *Annual TRB Meeting*. Washington, 1995.

[Wu00] N. WU: „Verkehr auf Schnellstraßen im Fundamentaldiagramm – Ein neues Modell und seine Anwendungen". In: *Straßenverkehrstechnik* 8 (2000).

[Zha09] X. ZHANG and S. CHEN: "Modeling of User's Route Choice Behavior in Adverse Weather". In: *Proceedings of 2009 IEEE International Conference on Grey Systems and Intelligent Services*. Nanjing, China, Nov. 10–12, 2009.

Corresponding author: Bernhard Heilmann, Austrian Institute of Technology, Department Mobility, A-1210 Vienna, Austria, phone: +43 50550 6338, e-mail: bernhard.heilmann@ait.ac.at

Method for the Organization of Daily Activity Chains

Domokos Esztergár-Kiss, Dénes Válóczi
Budapest University of Technology and Economics

Abstract

In the field of transportation the organization of daily activity chains has become more stressed, because the fast execution of the numerous tasks is a primary aspect. In order to reach high performance in the organization of the tasks, the attributes of the demand points, the transportation network and the external circumstances, as the changing traffic situation also have to be taken into account. A theoretical model was developed to organize and supervise the daily activity chains. Our aim is to improve the basically for logistic processes used TSP method and apply it for personal transportation purposes. The method offers a location based service, which results the optimal order of the tasks based on subjective parameters.

Keywords: daily activity chains, method development, TSP, flexible points, priorization of activities

1 Introduction

The ITS developments are widely spread in the field of passenger transportation. One direction of them is the LBS (= Location Based Services) technology, which uses the geographical data of a journey in order to utilize the demanded services of the passengers. The service is realized by the passengers' mobile devices using ITS, mobile internet and localization technology. With this extra information the passengers can plan their daily journeys in a more intelligent and optimized way.

The organization of the daily activity chains has been scrutinized in many articles [Hin12], [Tim03], [Mil03] and books [Tim05]. Organizing the chains some periodical repeated activity can be revealed (e.g. going to the office), which depends on the demographical [Ker07], on the spatial situation [Bul08] and on the personal characteristics of the user [Kan10]. More measurements were conducted in order to define the visited points, the average travel distance and time [Kam11], and in general the way of organizing the chains [Nij12], [Doh05]. Nevertheless only few articles were written in the topics related to the spatial and temporal

solutions [Doh06], to dynamic planning [Roo05], [Nij09], [Mar11] and to resolving possible conflicts [Aul08].

For sorting and ordering each activity the TSP (= Traveling Salesman Problem) method offers a solution [Rei94], [App07], which popular version is often called VPR (= Vehicle Routing Problem) [Tot02], [Gol08]. The solution for this problem was developed already 50 years ago, and since numerous versions were implemented. The basic problem is that an order has to be set among the points to be visited according to a specified aspect. This aspect could be travel distance, cost, number of transfers or the combination of these.

Basically the TSP method is used in logistics systems, but we propose an application for passenger transportation. In our case another constraint has to be defined, because the opening times of the shops and institutions have to be considered. Therefore our proposed algorithm is based on the TSP-TW (= TSP Time Window) version. Numerous articles deal with the problems and solutions of the TSP-TW method [Bal11], [Kos92], [Dum95], [Sav92], [Kol87], [Ghi11], [Das12]. The method has to be extended with flexible points, which are variable in time and space. The aim of the research would be to prove the benefits of this method compared to the basic TSP method.

2 The concept of the method

While establishing activity chains it is assumed that the passenger is already aware of the activities, which he/she would like to realize on the given day. The aim of the method is to set an order of the activities using the existing data (time, location, importance). The nowadays available methods (Fig. 1.), which we call basic TSP are based on activity points, which have to be explored and cost functions, which describe the values among the points.

We consider a TSP method flexible, if – according to the subjective demands of the travelers – some points can be arbitrarily replaced with another point of the same function. Therefore using the flexible TSP method a solution could exist, which would not exist using the basic TSP.

The idea of the predictive TSP can be explained as an extension of existing services , which can be derived from the demands. These latent demands of the passengers can be guessed from the demands of passengers with similar characteristics.

In this research we consider the flexible TSP, which includes the following steps:

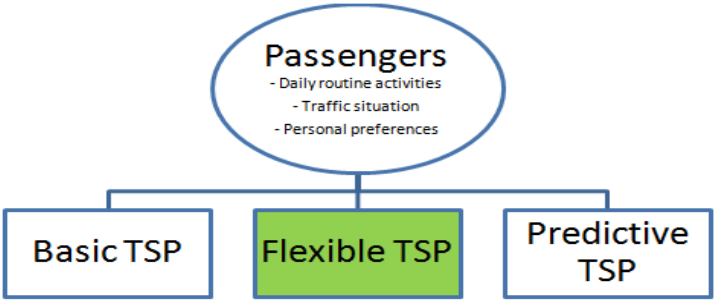


Figure 1: Types of TSP methods.

1. Definition of the daily activity chain:

- The list contains all the regular and non-regular activities of a passenger.
- The spatial and temporal parameters of the regular activities are usually fixed (e.g. school, workplace), while in many cases the non-regular activities are flexible.
- We assume that the passenger knows the activities of the certain day in advance and prepare the list of them, in which the time windows (TW) and processing time, which is the time spent at the points (TP), are also set.

2. Solving the basic TSP:

- The originally chosen points of the activity chain are the inputs for a TSP-TW method.
- The TSP-TW algorithm calculates an order of the points, which result could be the basic for a comparison with the proposed method.
- We generally assume that the basic problem is solvable, thus each point is reachable during the given TW (the TW-s are long enough and the TD-s are short enough).

3. Priorization of the activities:

- To each point a value is assigned, which represents its importance.
- The regular points of the activities get usually high priority, because they are spatially and temporally bounded, while the non-regular activities get lower priority.
- The subjective parameters are defined, all non-regular activities get a priority according to the personal demands and characteristics of the passenger.

4. Replacement of the flexible points:

- In the case of the flexible points the demanded service is reachable in more spatial places (e.g. instead of the point C, also in the point C'). Thus a new set of points could be installed, which is a new version of the activity plan, and the total travel cost could be reduced (Fig. 2.)
- The search for new points in the case of flexible demands is conducted according to the spatial distance with weighting, thus the closer is the new point to an existing point, the higher is its weight.
- Using the subjective parameters a further modification of the points is possible, but it has to be taken into account, when the passenger wants to perform the activity.
- If the basic problem would not be solvable (e.g. 2 points have the same TW, and the travel time is too long between them), it could be solved with the spatial and temporal replacement of the points (Fig. 2.).

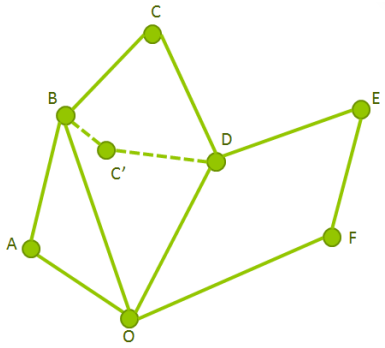


Figure 2: Presentation of the flexible points.

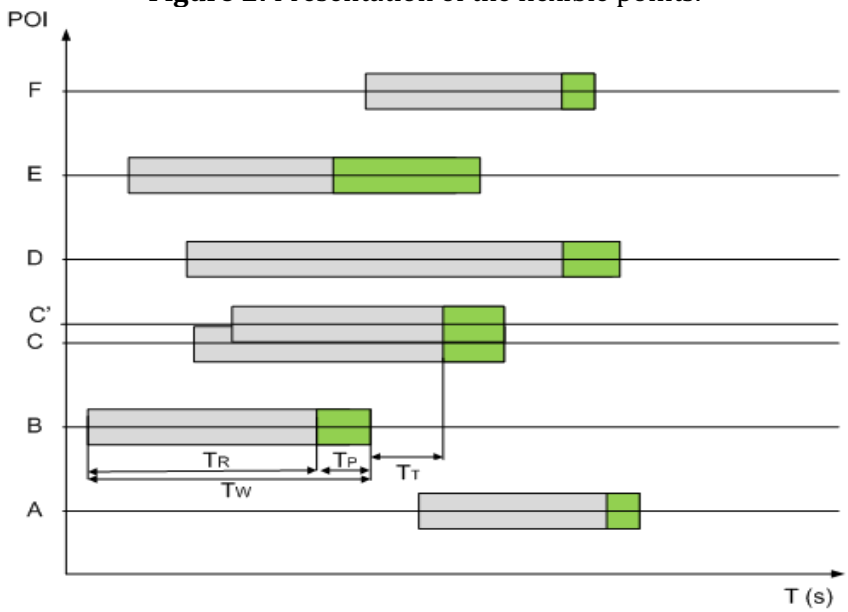


Figure 3: Temporal reachability of the points.

5. Optimization:

- For each version a TSP-TW is calculated.
- The version with the lowest total journey time (T) is chosen.
- The result of the basic TSP-TW is compared to the best version of the flexible TSP.

Concerning the points (Fig. 3.) the real time window (TR) of the arrival is defined by the original time windows (TW), which are usually the opening times of the shops and by the processing time (TP), which is the time needed for some operations executed (e.g. shopping). An average TP value (TP_a) can be defined, which can be modified to a minimal value (TP_m), if a delay occurs. The TP_m value is the time, which has to be minimal spent at the given point.

$$TR = TW - TP_a \tag{1}$$

By the flexible points if the next point (in the figure the point C) is not reachable during the travel time (TT) between the two points, then the algorithm searches for another point (C'), which is the closest to the prior point. Thus the needed activity can be performed at the new point (C').

The following constrains have to be fulfilled when using the model:

- each time window (TW) is at least as long as the processing time (TP),

$$TW \geq TP \tag{2}$$

- a point is reachable, if the real time window (TR) and the demanded time window of the passenger (TD) suits the processing time interval (TP),

- each (already replaced) point is reachable during the travel time (TT),

$$TT \geq TR_n - TR_m \tag{3}$$

- the total journey time (T) is the sum of all travel time (TT) and processing time (TP) and potentially waiting time (Twait), which can occur between the points.

$$T = \sum TT + \sum TP + \sum T_{wait} \tag{4}$$

3 First steps of the application's realization

The first step of the implementation is the elaboration of a data model (Fig. 4.), which is important to the construction of the daily activity chain. The model contains a Passenger table, where data about the passengers are stored. Some personal parameters belong to all passengers, which can be found in the Preferences table. These could denote preference of the transportation mode, disabilities or other factors. All preferences can be set in the interval of 1-5, which defines, how much the passenger demands the given service.

The certain points are contained in the POI table, which describes the name of the point, its address, opening hours and other information. These POI-s can be classified into types (e.g. food, sport, institutes and other categories), which is important because usually the passengers do not search for specific shops, but activity types. Using these categories finding the demanded activity becomes easier.

The most important table is the Activity, where the passengers can construct their daily activity chains. To each activity, which can be realized at the point (POI), belong time intervals. The duration, the time when passenger would demand the service (TD) and the processing time (TP), which is the duration of the service. The POI_ID field is optional, when filled, it is a fix point. Furthermore the priority should be filled, which denotes the importance of the activity. These can take the following values:

- 1: fix point, definitely has to be arranged on the certain day,

- 2: temporally fix, but spatially flexible,

- 3: temporally and spatially reduced flexible, namely the point has to be visited on the given day,

- 4: totally flexible, the point could be shifted to another day if necessary (e.g. if reaching the point would take more, than 1 hour and the schedule is too tight).

In the next step using the TSP-TW method the travel times can be calculated and optimized according to the activity chain. Then the order of the points to be visited can be defined. The exact elaboration and implementation of this algorithm is the next relevant step of the research.

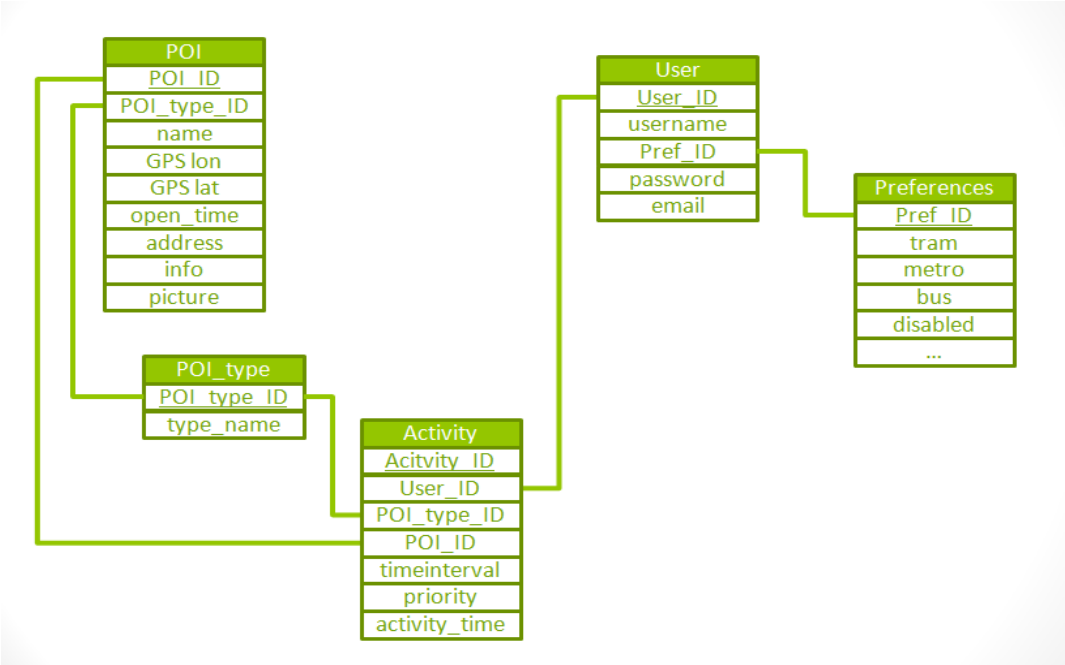


Figure 4: Data model and connections.

4 Further development

The theoretical model building has to be followed by the elaboration of practical part of the method and the analysis with VISUM simulation models. Still already some extensions and development directions emerged during the research process.

In our model we assumed that the passenger's start and end point is the same location. In the most cases it is valid, because usually the daily activity chains start from home and after finishing all activities the destination is home again. But in some cases the destination can be different, which requires another TSP method.

In the general case the cost function possesses fix values, thus among each points a fixed travel time (TT) is defined, but considering the actual traffic situation, the elements of the matrix could be changed in every hour. Using the same method the results of traffic jams and accidents could also be built in the model, which means the changing of the travel times.

The travel times could be defined according to the time tables of the public transportation and using real-time data the proper travel times could be assigned, which would contain the waiting times, the transfer times and the delays.

The cost matrix could be interpreted as a general resistance function, which takes into

account the travel times, the costs, the number of transfers and the personal preferences.

Enhancing the dynamics of the model the changes of the activity plan during the day could be taken into consideration (Fig. 5.). Compared to the original activity chain (Activity) some changes could occur, as appearing a new demand (and a new point) or a delay at a point (Changes). Thus the daily plan has to be re-planned and recalculated during the day (Algorithm). Using this method the real activity list could be followed and modeled (real Activity). Using the predictive TSP also the latent demands of the passengers could be served, and other services could be recommended for the certain passenger based on crowd sourcing data.

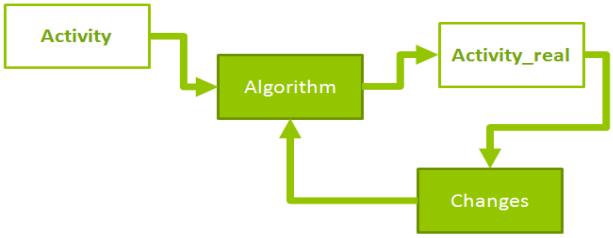


Figure 5: Dynamic modeling.

5 Summary

The organization of the passengers' daily activity chains is a complex ITS development field, which can be solved using the TSP-TW method. In the article we explored the problems of the organization and elaborated a method, which introduces flexible points and priorization. The highest priority means an important task, which definitely has to be arranged on the certain day. Lower priorities mean that the task can be even shifted to another day if necessary. Thus better results can be achieved than using the general TSP method. As the first step of the research a data model was defined, which contains data of the passengers, the points and the activities. Using this database an algorithm can be executed, which results in an optimal order of the points to be visited. The future development directions were also specified.

Acknowledgement

TÁMOP-4.2.2.C-11/1/KONV-2012-0012: "Smarter Transport" - IT for co-operative transport system - The Project is supported by the Hungarian Government and co-financed by the European Social Fund.

References

[App07] D. L. Applegate, R. E. Bixby, V. Chvátal, and W. J. Cook: *The Traveling Salesman Problem: A Computational Study*. Princeton: University Press, 2007. ISBN: 9780691129938.

[Aul08] J. Auld, A. K. Mohammadian, T. Sean, and S. T. Doherty: "Analysis of Activity Conflict

Resolution Strategies". In: *Transportation Research Record: Journal of the Transportation Research Board* 2054 (2008), pp. 10–19.

[Bal11] R. BALDACCI, A. MINGOZZI, and R. ROBERTI: "New State-Space Relaxations for Solving the Traveling Salesman Problem with Time Windows". In: *INFORMS Journal on Computing* 24.3 (2011), pp. 356–371.

[Bul08] R. N. BULIUNG, M. J. ROORDA, and T. K. REMMEL: "Exploring spatial variety in patterns of activity-travel behaviour: initial results from the Toronto travel-activity panel survey (TTAPS)". In: *Transportation* 35.6 (2008), pp. 697–722.

[Das12] S. DASH, O. GÜLNÜK, A. LODI, and A. TRAMONTANI: "A Time Bucket Formulation for the Traveling Salesman Problem with Time Windows". In: *INFORMS Journal on Computing* 24.1 (2012), pp. 132–147.

[Doh05] S. T. DOHERTY: "How far in advance are activities planned? Measurement challenges and analysis". In: *Transportation Research Record: Journal of the Transportation Research Board* 1926 (2005), pp. 40–49.

[Doh06] S. T. DOHERTY: "Should we abandon activity type analysis? Redefining activities by their salient attributes". In: *Transportation* 33.6 (2006), pp. 517–536.

[Dum95] Y. DUMAS, J. DESROSIERS, E. GELINAS, and M. M. SOMOLON: "An Optimal Algorithm for the Traveling Salesman Problem with Time Windows". In: *Operations Research* 43.2 (1995), pp. 367–371.

[Ghi11] G. GHIANI, E. MANNI, and B. W. THOMAS: "A Comparison of Anticipatory Algorithms for the Dynamic and Stochastic Traveling Salesman Problem". In: *Transportation Science* 46.3 (2011), pp. 374–387.

[Gol08] B. L. GOLDEN, S. RAGHAVAN, and E. A. WASIL: *The Vehicle Routing Problem: Latest Advances and New Challenges*. Springer, 2008, pp. 389–417. ISBN: 978-0-387-77778-8.

[Hin12] J. HINE, MD. KAMRUZZAMAN, and N. BLAIR: "Weekly activity-travel behaviour in rural Northern Ireland: differences by context and socio-demographic". In: *Transportation* 39.1 (2012), pp. 175–195.

[Kam11] M. KAMRUZZAMAN, J. HINE, B. GUNAY, and N. BLAIR: "Using GIS to visualise and evaluate student travel behavior". In: *Journal of Transport Geography* 19.1 (2011), pp. 13–32.

[Kan10] H. KANG and D. M. SCOT: "Exploring day-to-day variability in time use for household members". In: *Transportation Research Part A: Policiy and Practice* 44.8 (2010), pp. 609–619.

[Ker07] J. KERR, L. FRANK, J. F. SALLIS, and J. CHAPMAN: "Urban form correlates of pedestrian in youth: differences by gender, race-ethnicity and household attributes". In: *Transportation Research Part D* 12.3 (2007), pp. 177–182.

[Kol87] A. W. J. KOLEN, A. H. G. RINNON KAN, and H. W. J. M. TRIENEKENS: "Vehicle Routing with Time Windows". In: *Operations Research* 35.2 (1987), pp. 266–273.

[Kos92] Y. A. KOSKOSIDIS, W. B. POWELL, and M. M. SOMOLON: "An Optimization-Based Heuristic for Vehicle Routing and Scheduling with Soft Time Window Constraints". In: *Transportation Science* 26.2 (1992), pp. 69–85.

[Mar11] F. MARKI, D. CHARYPAR, and K. W. AXHAUSEN: "Continuous activity planning for a continuous traffic simulation". In: *Transportation Research Record: Journal of the Transportation Research Board* 2230 (2011), pp. 29–37.

[Mil03] E. J. MILLER and M. J. ROORDA: "Prototype Model of Household Activity-Travel Scheduling". In: *Transportation Research Record: Journal of the Transportation Research Board* 1831 (2003), pp. 114–121.

[Nij09] E. W. L. NIJLAND, T. A. ARENTZE, A. W. J. BORGERS, and H. J. P. TIMMERMANS: "Individuals' activity – travel rescheduling behaviour: experiment and model-based analysis". In: *Environment and Planning A* 41.6 (2009), pp. 1511–1522.

[Nij12] L. NIJLAND, T. ARENTZE, and H. TIMMERMANS: "Incorporating planned activities and events in a dynamic multi-day activity agenda generator". In: *Transportation* 39.4 (2012), pp. 791–806.

[Rei94] G. REINELT: *The traveling salesman: computational solutions for TSP applications*. Berlin, Heidelberg: Springer, 1994. ISBN: 3-540-58334-3.

[Roo05] M. J. ROORDA and E. J. MILLER: *Strategies for Resolving Activity Scheduling Conflicts: An Empirical Analysis – Progress in Activity-Based Analysis*. Oxford: Elsevier, 2005. pp. 203-222. ISBN: 0080445810.

[Sav92] M. W. P. SAVELSBERG: "The Vehicle Routing Problem with Time Windows: Minimizing Route Duration". In: *INFORMS Journal on Computing* 4.2 (1992), pp. 146–154.

[Tim03] H. TIMMERMANS, P. VAN DER WAERDEN, M. ALVES, J. POLAK, S. ELLIS, A. S. HARVEY, S. KUROSE, and R. ZANDEER: "Spatial context and the complexity of daily travel patterns: an international comparison". In: *Journal of Transport Geography* 11.1 (2003), pp. 37–46.

[Tim05] H. TIMMERMANS: *Progress in activity-based analysis.* Elsevier Science, 2005. ISBN: 9780080445816

[Tot02] P. TOTH and D. VIGO: *The Vehicle Routing Problem*. SIAM, 2002, pp. 157-186. ISBN: 978-0-898715-79-8.

Corresponding author: Domokos Esztergár-Kiss, Budapest University of Technology and Economics, Department of Transport Technology end Economics, 1111 Muegyetem rkp. 3-11., Budapest, Hungary, +36 1 463 1029, esztergar@kku.bme.hu

Minimization of Vehicle Stops by an Early Termination of Green Times in Traffic-Light Controlled Road Networks

Kathleen Tischler, Stefan Lämmer
Technische Universität Dresden

Abstract

Traffic lights are typically operated with cycle times and green splits being optimized for average expected demands. Unexpected demand fluctuations on shorter time scales require reactive adjustments to the green times. A commonly applied technique is to terminate green times within a scheduled period if the time interval to the next arriving vehicle exceeds a pre-defined value. However, this so-called vehicle interval method does not consider arrivals on subsequent stages.

The present paper proposes to anticipate the number of vehicle stops in the current as well as in the next stage as a function of possible termination time points. A stage transition is initiated when the total number of stops is at a minimum. The window of possible termination time points can be defined for each stage separately. Additionally, synchronization time points guarantee fixed cycles and offsets for coordinated movements between intersections. As a model based alternative to the vehicle interval heuristics, the proposed method is compatible with cycle-based control strategies. Comparative simulations of a coordinated arterial indicate highly significant reductions of both, the number of vehicle stops and delay times.

Keywords: traffic light control, urban traffic, coordination, vehicle stops

1 Introduction

Vehicular traffic has shaped urban mobility throughout the last decades. As the functionality of the underlying infrastructure crucially depends on how road intersections are operated, much effort has been put into the investigation of traffic light control strategies that increase throughput and reduce delays. The first advancements have been made by fixed-time strategies that were optimized offline based on historical traffic data, for example, TRANSYT [Rob69]. In contrast, online control strategies aim to adapt to steadily changing traffic conditions by a permanent feedback of real-time detector data. The probably most

famous examples are SCOOT [Hun81] and SCATS [Sim80]. A broad overview of these and other approaches is given in the review articles [Hou01; Pap03; Pap07].

A particular difficulty in the optimization of signal timings is the anticipation of traffic demands. As both the adaption and re-optimization of control parameters take place on time scales of several 10 minutes even with modern technology, assumptions have to be made. Some optimization methods assume periodic arrival patterns, such as OPAC [Gar83] or RHODES [Mir01] whereas others assume uniform or stochastic arrivals, such as TUC [Dia03] or MOVA [Vin88] which is based on Miller's algorithm [Mil63]. Although MOVA detects arrivals for the current stage as our proposed method, it further anticipates a uniform traffic flow rate.

In order to react on short-term demand fluctuations, these strategies generally incorporate traffic-responsive elements. Cycle-based techniques commonly use the vehicle interval method. It terminates green times if (besides other conditions) the time gap to the next arriving vehicle is larger than a specified threshold. As large time gaps usually imply that an initial queue has been resolved and the majority of arrivals have been served, an early transition to the next stage will in most cases be more beneficial in terms of total vehicle delays and stops than the continuing with the present one. This heuristics fails, however, in two distinct situations: (a) A significant number of vehicles arrive after the critical gap was exceeded so that all of them have to stop at red. (b) A few stray vehicles postpone the transition to the next stage where an earlier green start would have prevented a larger vehicle platoon from being stopped.

With the purpose to overcome these problems, this paper presents an alternative way to identify suitable termination time points. The method anticipates the number of vehicle stops caused by an early termination of the current green times. Then, they are compared against the expected number of stops that can be avoided by an earlier green start of the next stage. The chosen transition time point is determined by mimimizing the total number of stops under the conditions that (i) the termination time is within a specified time window, (ii) all initial queues have been cleared. As the proposed method optimizes for approaching vehicles within a detection horizon, it can be classified as a rolling horizon technique.

The paper is organized as follows. Chapter 2 introduces the concept of stop anticipation, chapter 3 formulates the control strategy, and chapter 4 quantifies the potential of stop minimization by the simulation of a coordinated arterial. The paper concludes with a discussion of the most relevant results.

2 Methodology

We consider a traffic light controlled intersection with a given cycle-based signal timing. The cycle is divided into stages s, in which all associated signal groups i show green simultaneously and control traffic movements on one or more lanes l. Since the green times of the signal groups often do not start (or end) at the same time point, a stage is defined as the period in which each associated signal group displays green. Hence, the duration of a stage

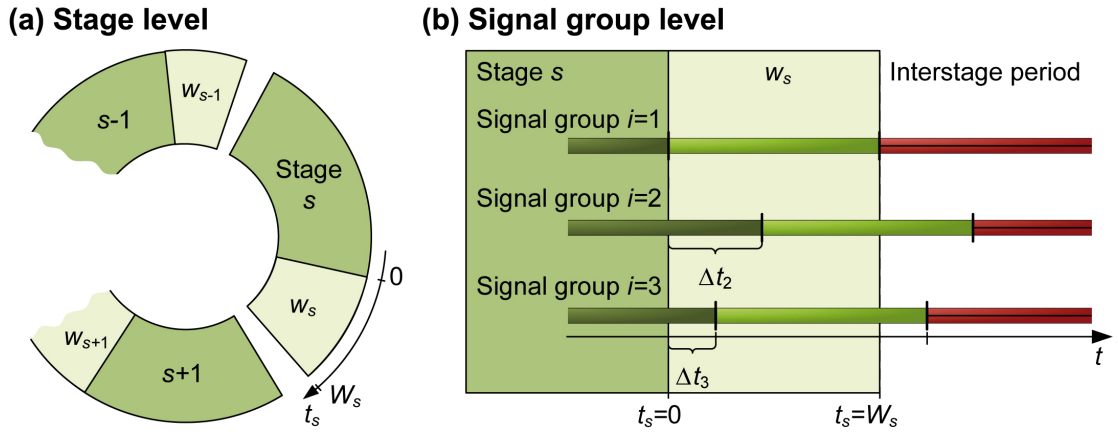


Figure 1: (a) The cycle of a given signal timing plan is divided into stages s (green), and interstage timing periods (blank). Possible stage termination time points t_s are defined by the window w_s (light green). (b) The individual green end for an associated signal group i is shifted by Δt_i and lies within $[\Delta t_i, W_s + \Delta t_i]$ (light green bar).

is generally shorter than the green times of the individual signal groups. Interstage timing periods separate the stages.

2.1 Stage Transitions

Each stage s might be associated with a window w_s of possible termination time points. A window is assumed to cover the last W_s seconds of the correspondig stage period as Fig. 1 indicates. As soon as such a window is reached and all the conditions to be developed below are fulfilled, the controller immediately proceeds with the subsequent interstage timing period and thereby initiates the transition to the next stage. An early termination of the previous stage causes all signal groups of the following stage $s + 1$ to start its green times and its asociated window earlier by the same amount of time. The window end can be forwarded as well, or it remains fixed in the cycle. If we decide for fixed window ends, the previously saved green times are applied to extend the window.

In order to compensate for an undesired time drift, and to guarantee fixed cycle and offset times as required for coordination, at least one synchronization time point per cycle has to be included. Whenever the potentially advanced controller's time coincides with such a synchronization time point, it pauses until it is synchronized with the actual time again. Synchronization time points are to be placed within a stage period before the start of a window. Hence, the windows of those stages remain fixed. Typically, all stages that control coordinated traffic flows should be synchronized.

2.2 Termination Time Points

The termination time point of stage s relative to the associated window w_s is denoted by t_s. The value of $t_s = 0$ ($t_s = W_s$) corresponds to the start (end) of an associated window w_s,

such that $t_s \in [0, W_s]$, see Fig. 1 (a). Note that, in terms of absolute time, the green times of the associated signal groups $i \in I_s$ do not end simultaneously at the termination of stage s. Since the green ends are constantly shifted by a time interval Δt_i, which is set according to the signal timing plan, some signal groups might end within the subsequent interstage timing period, as shown in Fig. 1 (b). Therefore, it is necessary to consider individual termination time points $t_i = t_s + \Delta t_i$. The same applies to the signal groups $j \in I_{s+1}$ of the next stage $s + 1$. Their individual starts t_j are linked to t_s by $t_j = t_s + \Delta t_j$, where Δt_j are generally scheduled to be larger than Δt_i because of the intergreen times. This notation will be used in the following to anticipate the number of vehicle stops that different termination time points will produce or avoid.

2.3 Arrival Detection

At least one inflow detector at each lane l of signal group i measures the crossing time points of the approaching vehicles sufficiently far ahead. A stop line detector per lane detects vehicles that pass the stop line. The free-flow travel time from the inflow detector to the stop line is assumed to be known and should be larger than the previous interstage timing period plus the time required to clear an initial queue. For example, given them to be 5 s and 8 s, the free-flow travel time should be larger than 13 s. At a velocity of 50 km/h, it requires an inflow detector distance of at least 180 m. To satisfy the minimum desired distance if intersections are closely spaced, stop line detectors of upstream intersections could be utilized. Possible additional inflow detectors allow to correct the arrival prognosis in case of variable velocities and lane changes. Under the assumption that detector failures are neglectable, this setup allows an estimation of queue lengths, for example, by applying the section-based-model [Hel03; Tre13]. Furthermore, it allows an anticipation of the green time required to clear a queue under the presence of irregular arrivals [Lam07].

Each detection impulse at an inflow detector indicates that a new vehicle $v \in V_{s,l,i}$ is about to arrive at the associated lane l. Under free traffic conditions, i. e. if the regarded vehicle is neither delayed nor stopped, it will pass the stop line later than the detector impulse occured, namely after the according free-flow travel time. Its earliest possible arrival time point at the stop line of lane l of signal group i is anticipated, and denoted by $t^{\text{arr}}_{v,l,i}$. In order to account for the number of vehicle stops, we need to evaluate for all approaching vehicles whether the traffic light shows green and whether the queue is cleared at the time when they could arrive at the stop line. The initial queue length can be anticipated by first checking remaining queues at the end of a green time from the balance of the time-shifted inflow and the stop line detector count. Additionally, we count the inflow detection impulses for all vehicles that will arrive at the stop line during red and during the queue clearing process.

2.4 Vehicle Stops in the Current Stage

We consider a state in which all signal groups i of stage s show green, first, discharging initial queues and then serving arrivals. Since the control strategy as developed below will

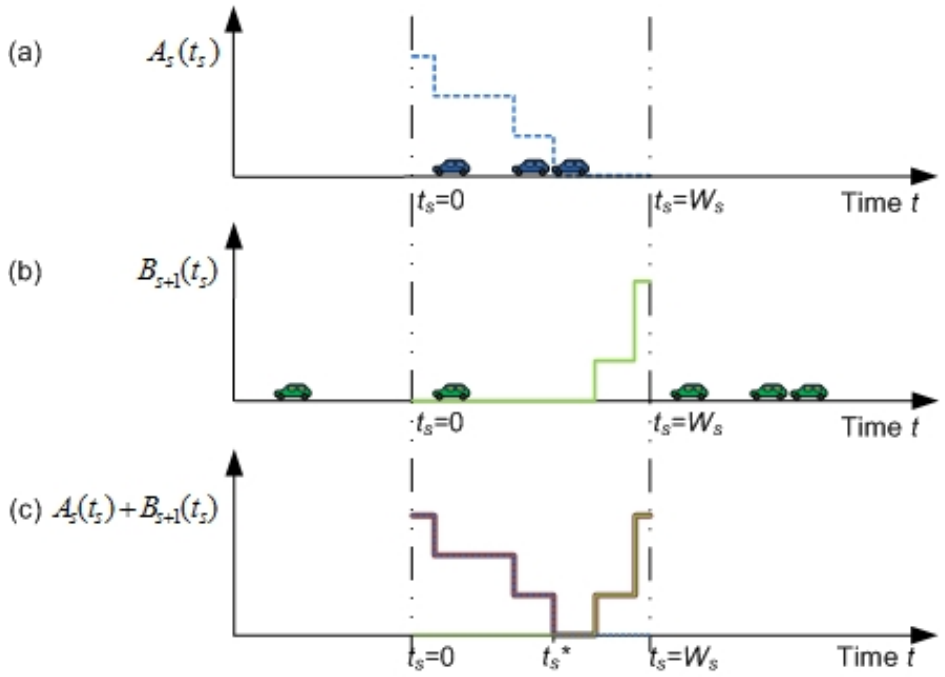


Figure 2: Anticipated number of vehicle stops as a function of the termination time (a) in the current stage, (b) in the next stage, and (c) in total.

not optimize for a termination before all queues have been cleared, it is not neccessary to account for queue-related stops. We will also not account for vehicles that arrive after the window end W_s, as they always have to stop for a red light independent of whether the stage was terminated earlier or not. Therefore, the total number of vehicle stops $A_s(t_s)$ that result from an early termination of the present green times can be written as a function of the termination time point t_s.

$$A_s(t_s) = \sum_{i,l,v} \alpha_{i,l,v}(t_s) \quad \text{with} \quad \alpha_{i,l,v}(t_s) = \begin{cases} 1, & \text{if } (t_s + \Delta t_i < t^{\text{arr}}_{v,l,i} < W_s + \Delta t_i) \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

If vehicles $v \in V_{s,l,i}$ of lane $l \in L_{s,i}$ arrive later than the associated signal group $i \in I_s$ terminated, they have to stop. Hereby, the indicator function $\alpha_{i,l,v}(t_s)$ is used to select those vehicles. The function $A_s(t_s)$ describes a stepwise decreasing curve that reaches zero at the window end $t_s = W_s$. Any earlier termination time point $t_s < W_s$ might result in a positive number of stops dependent on how many vehicles arrive between t_s and W_s, see Fig. 2 (a).

2.5 Vehicle Stops in the Next Stage

After the current stage s is to terminate at time point t_s, the green times of the signal groups j of the next stage $s + 1$ will start accordingly sooner. Consequently, initial queues will be cleared earlier and further arrivals are potentially prevented from being stopped. Since the queued vehicles can only leave one after the other, the earliest time at which vehicle

$v \in V_{s+1,l,j}$ on lane l can pass the stop line is the time $t_j = t_s + \Delta t_j$ at which signal group j has turned to green, plus the time τ_v that all other vehicles in front of v needed to depart. Here, we count from the first stopped vehicle $v = 1$ in the initial queue. Hence, τ_v can be estimated as the number of vehicles in front of v divided by the saturation flow rate s_l of the associated lane: $\tau_v = (v-1)/s_l$. Accordingly, vehicle v will have to stop if it arrives before $t_j + \tau_v$, i. e. before all previous vehicles have departed on green. This leads to the following expression for the number of stops $B_{s+1}(t_s)$ that occur on lane $l \in L_{s+1,j}$ of signal group $j \in I_{s+1}$ in the next stage $s+1$ as a function of the termination time point t_s.

$$B_{s+1}(t_s) = \sum_{j,l,v} \beta_{j,l,v}(t_s) \quad \text{with} \quad \beta_{j,l,v}(t_s) = \begin{cases} 1, & \text{if } (\Delta t_j + \tau_v < t^{\text{arr}}_{v,l,j} < t_s + \Delta t_j + \tau_v) \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

The indicator function $\beta_{j,l,v}(t_s)$ accounts for only those stops that can be avoided if the previous stage terminates at t_s. Since the earliest possible green start of signal group j is $t_j = 0 + \Delta t_j$, the above formula does not capture vehicles with arrival time points earlier than $\Delta t_j + \tau_v$, i. e. it ignores stops that occur in any case (contrary to the estimation of τ_v). Therefore, the function $B_{s+1}(t_s)$ is zero at $t_s = 0$ and increases stepwise from there on.

Note the strong sensitivity of $B_{s+1}(t_s)$ on arriving vehicle platoons. Every vehicle v will extend the τ_v-value of its successor, denoted by τ_{v+1}. Hereby, v enlarges the lower bound $\Delta t_j + \tau_{v+1}$ as well as the upper bound $t_s + \Delta t_j + \tau_{v+1}$ of arrival time points and the scope of vehicles which $\beta_{j,l,v}(t_s)$ accounts for. In consequence, a platoon of n closely following vehicles will cause $B_{s+1}(t_s)$ to jump by the same amount n as soon as the termination time point t_s causes the platoon to stop, see Fig. 2 (b). Another implication of this oberservation is that $B_{s+1}(t_s)$ is not strictly limited. It can, unlike $A_s(t_s)$, be larger than the number of vehicles that arrive within the associated window w_s.

3 Control Strategy

The following control strategy is being applied: Given the window w_s of stage s has started, i. e. $0 \le t_s < W_s$, and all vehicle queues of the associated signal groups have been cleared. Then, the current stage s is terminated, and the transition to the subsequent stage $s+1$ is initiated as soon as the total number of anticipated stops

$$A_s(t) + B_{s+1}(t) \tag{3}$$

over all remaining termination time points $t_s \le t < W_s$ is at the minimum for the current time point t_s.

In case of multiple minima, i. e. if several different termination time points lead to the same minimum number of stops, the above strategy decides for the first time point. In particular, if the minimum ranges over a continuous time interval due to larger arrival headways, this strategy terminates the current green times right after the last vehicle has passed the stop

Table 1: Demand and Flow relations.

Approach	Origin traffic demand	Right turns	Through	Left turns
West (I1, I3, I5)	840 veh/h (I1)	5 %	95 %	–
East (I1, I3, I5)	640 veh/h (I5)	10 %	90 %	–
East/West (I2, I4)		10 %	80 %	10 %
North (I1-I5)	250 veh/h (I4: 200 veh/h)	30 %	50 %	20 %
South (I1-I5)	250 veh/h (I2: 200 veh/h)	40 %	30 %	30 %

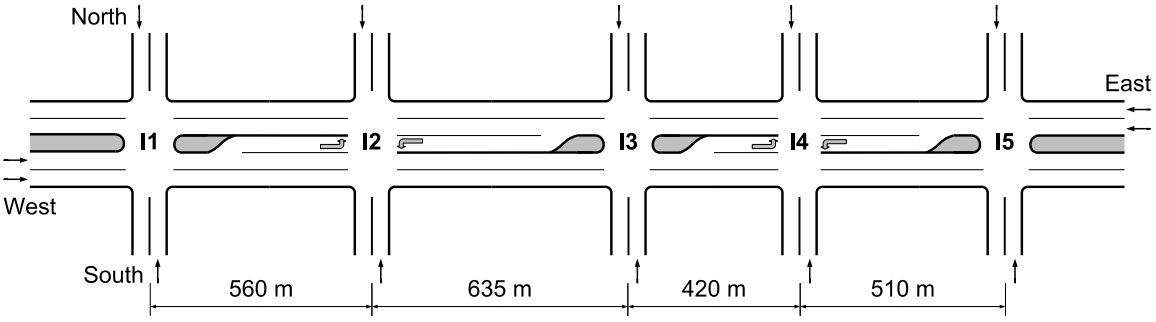


Figure 3: Arterial with five irregularly placed intersections.

line and turns on the green lights of the next stage sooner.

4 Simulation Results

The above control strategy, referred to as stop-minimization, has been implemented in Java and integrated to the microscopic simulation tool VISSIM using the COM-interface. We simulated an artificial arterial consisting of five intersections, as depicted in Fig. 3. Based on a set of fixed-time signal timing plans that was designed and optimized according to the Traffic Signal Timing Manual [Tra08], we implemented the fixed-time, the vehicle interval, and the proposed stop-minimization control at unsaturated conditions.

We applied inflow rates and turning relations as given in Tab. 1. Arrivals in the network were Poisson distributed. The signal timing was designed to propagate a green wave eastwards at a velocity of 50 km/h. Based on the most saturated intersection I4, a common cycle length of 90 s, the following green splits, and four stages were chosen: (1) the coordinated bidirectional arterial with through movements and right turns (split: 42 %), (2) main street left turns (12 %), (3) all northern movements (22 %), and (4) all southern movements (23 %). The other intersections give 42 % of the cycle to the coordinated main street as well. For those without main street left turns, the spared green times were proportionally assigned to the north and south approaches. In order to keep the saturation of green time at a level below 75%, the inflow at I2 south and I4 north was reduced (see Tab. 1).

Table 2: Average stops and delay times per vehicle based on 40 simulation runs of one hour, each. The relative savings refer to the fixed-time control. Corresponding t-values indicate a high significance ($p < 10^{-9}$).

Strategy	Stops	t-value	Delay	t-value
Fixed-time control	1,410	–	53,43 s	–
Vehicle interval control	1,356 (-3,9 %)	20,28	51,52 s (-3,6 %)	21,32
Stop-minimization	1,318 (-6,5 %)	36,96	49,71 s (-7,0 %)	39,75

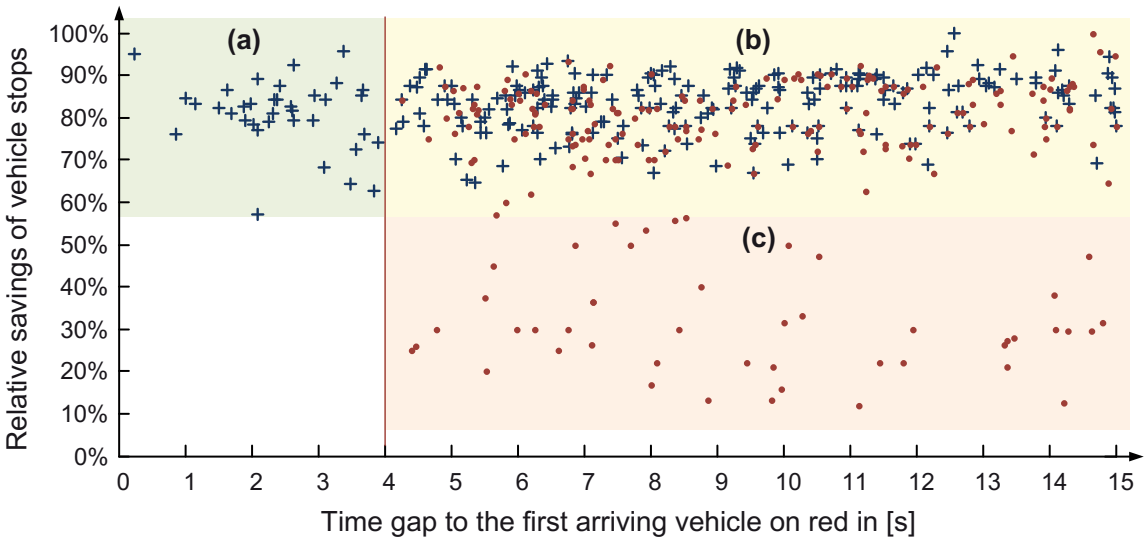


Figure 4: To compare the performance of the stop-minimization (crosses) and the vehicle interval control (dots), the control parameters of each were measured at the stage termination point. Vertical axis: The relative stop savings for both stages as ratio of undelayed vehicles to total arrivals that could have been stopped. Horizontal axis: The time gap from the green end to the next arriving vehicle at the stop line.

For the stop-minimization and the vehicle interval control, equivalent windows were allocated within the so specified green times. The window starts were defined: (a) for the non-coordinated stages immediately after a minimum green of 5 s, (b) for the coordinated stage after the green time that is necessary to serve the average expected number of vehicles at maximum flow rate. For the vehicle interval control, we chose a critical time gap of 4 s as Akçelik [Akc99] suggests in his study on queue discharge characteristics. For both strategies, saved green times from early terminations were allowed to extend further windows of the next stages up to the coordinated stage, in which a synchronisation time point was assigned to.

By setting the fixed-time control as the reference, the other two strategies performed significantly better (see Tab. 2). The proposed stop-minimization even outperfomed the vehicle interval control. Most termination decisions of the two strategies had the same effect

on stop savings as Fig. 4 (b) illustrates. Whereas the vehicle interval method was designed to wait for gaps above 4 s, the stop-minimization occasionally accepts smaller gaps in cases where many stops are prevented (see Fig. 4 (a)). Furthermore, the red dots in Fig. 4 (c) indicate that the vehicle interval control fails to save stops by waiting for large enough time gaps.

5 Conclusion

We proposed a method to enhance a coordinated traffic light control by traffic-responsive elements. In order to miminize vehicle stops, a stage is allowed to be terminated between a minimum and a maximum green time. The coordination based on common cycle length and fixed offsets was maintained due to (i) introducing synchronization time points and (ii) applying the concept of windows.

This concept is compatible with any cycle-based control strategy that is arranged in stages. Furthermore, it appears to be a significantly better performing alternative compared to the heuristic vehicle interval control in terms of vehicle stops and delay times. Since the proposed model-based approach considers all arrivals within the available prognosis horizon, and as it also accounts for arrivals in subsequent stages, it is more anticipative than the vehicle interval control. This might be the reason for the improved adjustment to short-term traffic fluctuations.

The way how the chosen termination time points influence the stages after the next is not explicitly evaluated. If a stage is terminated earlier, it can induce more vehicles waiting at the next start of the same stage, and this may postpone its termination. As well, the effect on the next intersection, for example, by an earlier starting platoon, is disregarded. The results of the simulation indicate that in most cases the controller of that intersection is able to react on the time changed arrivals in the same way. While the stop-minimization concept operates on a limited temporal as well as spatial horizon, it seems still able to adapt to global requirements. More broadly, research is needed to validate the effects in a more complex and realistic simulation scenario, for example, by implementing a real-world traffic network.

Further investigations have to be made on how to overcome limitations of vehicle detection. So far, a sufficient prognosis horizon has been assumed to detect every vehicle that is affected by an early termination. However, especially long queues in the next stage would require extensive horizons and thus impracticable large detector distances. Short lanes are not yet considered. Besides additional inflow detectors they need an advanced anticipation of how many vehicles will enter that lane. An additional objective is to enhance the control for oversaturated traffic conditions, for example, by extending the strategy with a capacity maximizing mode. Then, green times could be adjusted within the windows to increase departure flow rates.

Acknowledgments

The authors thank the DFG (German Research Foundation) for partial financial support of this research. Thanks also to Martin Treiber for useful discussions.

References

[Akc99] R. AKCELIK, M. BESLEY, and R. ROPER: *Fundamental relationships for traffic flows at signalised intersections*. ARR 340. ARRB Transport Research, 1999.

[Dia03] C. DIAKAKI, V. DINOPOULOU, K. ABOUDOLAS, M. PAPAGEORGIOU, E. BEN-SHABAT, E. SEIDER, and A. LEIBOV: "Extensions and new applications of the traffic signal control strategy TUC". In: *Transportation Research Board* 1856 (2003), pp. 202–211.

[Gar83] H. GARTNER: "OPAC: A demand-responsive strategy for traffic signal control". In: *Transportation Research Record* 906 (1983), pp. 75–84.

[Hel03] D. HELBING: "A section-based queueing-theoretical traffic model for congestion and travel time analysis in networks". In: *Journal of Physics A* 36 (2003), pp. L593–L598.

[Hou01] N. B. HOUNSELL and M. MCDONALD: "Urban network traffic control". In: *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering* 215.4 (2001), pp. 325–334.

[Hun81] P. HUNT, D. ROBERTSON, R. BRETHERTON, and R. WINTON: *SCOOT - a traffic responsive method of coordinating signals*. TRRL Report 1014. Crowthorne, Berkshire, UK: Transport and Road Research Laboratory, 1981.

[Lam07] S. LÄMMER, R. DONNER, and D. HELBING: "Anticipative control of switched queueing systems". In: *The European Physical Journal B* 63.3 (2007), pp. 341–347.

[Mil63] A. J. MILLER: "A computer-controlled system for traffic networks." In: *Proc. 2nd Intern. Symposioum on the Theory of Road Traffic Flow*. London, 1963, pp. 200–220.

[Mir01] P. MIRCHANDANI and L. HEAD: "A real-time traffic signal control system: architecture, algorithms, and analysis". In: *Transportation Research Part C: Emerging Technologies* 9.6 (2001), pp. 415–432.

[Pap03] M. PAPAGEORGIOU, C. DIAKAKI, V. DINOPOULOU, A. KOTSIALOS, and Y. WANG: "Review of road traffic control strategies". In: *Proceedings of the IEEE* 91.12 (2003), pp. 2043–2067.

[Pap07] M. PAPAGEORGIOU, M. BEN-AKIVA, J. BOTTOM, P. BOVY, S. HOOGENDOORN, N. HOUNSELL, A. KOTSIALOS, and M. MCDONALD: "ITS and Traffic Management". In: *Handbook in OR & MS*. Ed. by C. BARNHART and G. LAPORTE. Vol. 14. Elsevier, 2007, pp. 715–774.

[Rob69] D. I. ROBERTSON: *TRANSYT: a traffic network study tool*. Tech. rep. LR 253. Crowthorne, UK: Transport and Road Research Laboratory, 1969.

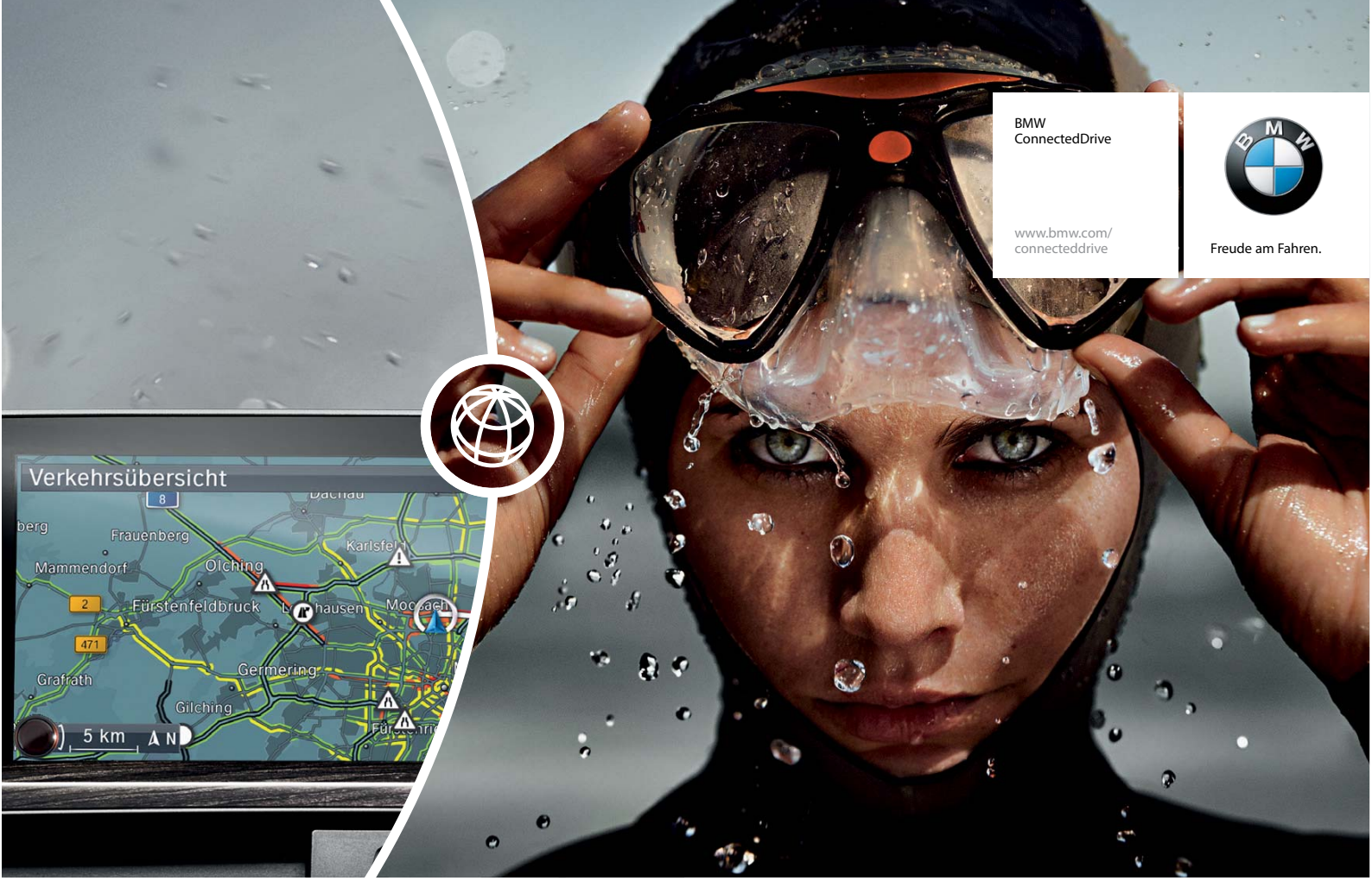
[Sim80] A. G. SIMS and K. DOBINSON: "The Sydney coordinated adaptive traffic (SCAT) system philosophy and benefits". In: *Vehicular Technology, IEEE Transactions on* 29.2 (1980), pp. 130–137.

[Tra08] U. D. of TRANSPORTATION: *Traffic Signal Timing Manual*. U.S. Department of Transportation. 2008.

[Tre13] M. TREIBER and A. KESTING: *Traffic Flow Dynamics – Data, Models and Simulation*. Springer, 2013.

[Vin88] R. VINCENT and J. PEIRCE: *'MOVA': Traffic responsive, self-optimising signal control for isolated intersections*. Research Report RR 170. Crowthorne, UK: Transport and Road Research Laboratory, 1988.

Corresponding author: Kathleen Tischler, Technische Universität Dresden, "Friedrich List" Faculty of Transportation and Traffic Sciences, 01062 Dresden, Germany, phone: +49 351 463 838, e-mail: kathleen.tischler@tu-dresden.de



BMW
ConnectedDrive

www.bmw.com/
connecteddrive

Freude am Fahren.

SCHNELLER AM ZIEL.

Vorankommen statt Stillstand. Präzise Informationen zur aktuellen Verkehrslage: mit Real Time Traffic Information haben Sie den Verkehr im Blick. Rush Hour, Baustelle oder Ferienstau – egal, wie viel los ist – Ihr BMW kennt den schnellsten Weg und sorgt so für eine entspannte Anfahrt. Verbringen Sie Ihre wertvolle Zeit dort, wo es Sie hinzieht, ob zu Ihrer Familie, zu Freunden oder an Ihren Lieblings-Spot.

REAL TIME TRAFFIC INFORMATION.

BMW ConnectedDrive
Vernetzt, um frei zu sein.

Using Nanoscopic Simulations to validate the Benefit of Advanced Driver Assistance Systems in complex Traffic Scenarios

Torsten Schubert, Mario Krumnow, Bernard Bäker, Jürgen Krimmling
Technische Universität Dresden

Abstract

The increasing traffic demand due to growing urbanisation results in higher emission density in urban regions. The main goals of today's research and development are leading to different systems and topics for more energy-efficient technologies in powertrains and intelligent driver assistance systems. Furthermore traffic actuated traffic lights are state of the art to optimize the traffic flow. The reciprocal effects among these improvements can be evaluated by a new simulation approach introduced in the present work.

This publication deals with an interface between detailed nanoscopic vehicle simulation with MATLAB/Simulink and traffic flow simulation with SUMO. The integration of the traffic simulation into MATLAB is described. To demonstrate the benefit of this simulation environment a traffic light assistance system is implemented in MATLAB and the results of the combined simulation are shown. The impact of different traffic situations on the global energy consumption can be determined. With this simulation (framework) different parameters like speed limits, traffic light control, number of lanes, and density of traffic flow can be evaluated.

Keywords: SUMO, MATLAB, Simulink, ADAS, traffic light assistance, efficiency, traffic simulation, urban, nanoscopic, microscopic, TraCI, TraaS

1 Introduction

Increasing traffic volume and growing urbanisation result in higher emission in urban regions. So the main goals of today's research and development are on different systems and topics for energy-efficient technologies in engines, powertrains, intelligent driver assistance systems and the infrastructure to provide higher traffic flow. Hence, increasing the efficiency of the control system 'driver-vehicle-traffic' is indispensable. According to [Dor04], to achieve this increase three opportunities and feasible measures exist:

- The infrastructure: e.g. intelligent traffic light control systems.

- The vehicle: e.g. control strategies for hybrid powertrains.
- The driver and his driving style: advanced driver assistance systems (ADAS).

Simulation of these systems is used to ensure the fulfilment of related quality, security and cost objectives prior to the implementation to the real environment. These systems are strongly interconnected as they mutually influence each other. In fact, most traffic scenarios have numerous aspects that influence a driver and the energy consumption of a vehicle, e.g. speed limits, traffic lights and other road users. These influences have to be identified and analysed using a tool for traffic simulation in combination with detailed vehicle simulation.

2 Current Approaches for detailed Vehicle Simulation in traffic scenarios

For simulation of vehicle dynamics [Tem12], Control strategies for hybrid and electric vehicles [Kut11] and Advanced Driver Assistance Systems (ADAS) [Sch11] different tools can be used. The most preferred tool for that purpose is *MATLAB*®. It is very powerful for numerical computation, visualization, optimization and programming [MAT13]. However, in most cases only generated or recorded driving cycles are used for simulation. Methods are needed to analyse the usability of ADAS and their effects in a traffic simulation. First approaches in that field are provided in [Bia05], [Hab11] and [Ved13]. But they are only suitable for very simple scenarios.

For the simulation of traffic scenarios numerous commercial software suites are available e.g. VISSIM, PARAMICS, AIMSUN as well as free software like SUMO (DLR) or MATSim [Kok11]. These tools are mainly used to determine incidents or congestions in traffic networks. There has been a trend in recent years towards approaches that divides the vehicle into single parts to simulate each component separately. On the one side some commercial tools are designed for the simulation of vehicle dynamics like PELOPS [FKA13] or CarMaker [IPG13]. On the other side there are tools to compute the emission of vehicles available e.g. Paramics Software Suite [Qua13]. In order to ensure individual adaptation to the purpose on the simulation of nanoscopic *Simulink* models the Open Source microscopic traffic simulation *SUMO* [Ber11] is a good choice. *SUMO* is very fast, powerful and available for different operation systems (MS Windows, Mac OS and Linux). Another advantage is the developer community which provides a software update nearly every day. Further more than five different car-following-models are already implemented which can be adjusted in a configuration file. At least the most important fact is the possibility to interact with the simulation using an I/O-interface during runtime [Ber11]. That interface is frequently used and extended in numerous research topics e.g. [Kru13c], [Som11].

SUMO already contains a model for emissions and fuel consumption [Ber11] which is based on the HBEFA database [INF13]. But the current model lacks of detail, e.g. no regard to the selected gear and the dependent RPM which leads to an inaccurate computation of emission values. There are some approaches for the integration of more detailed electric vehicle energy models in *SUMO* [Mai11], [Kur13], to get a first impression for the need of higher precision values of the simulation results.

Also different approaches using *MATLAB* [MAT13] for simulation of traffic scenarios exist. In [Mac13] a coupled simulation of *SUMO* and *MATLAB* using an interface called TraSMAPI is described. The Information from *SUMO* is captured and processed in *MATLAB*/*Simulink*. According to this, it is evident that they lack of a feedback from *MATLAB* to *SUMO*. It is not possible to manipulate parts of the simulation for the evaluation of ADAS. Thus is necessary to give *MATLAB* the opportunity to interact with *SUMO* bidirectionally.

3 Extend detailed vehicle dynamic models with microscopic traffic simulations

Today it is achievable to have huge microscopic simulations which handle every vehicle (agent) separately in every simulation time step. In general there are two models which represent the specific driver behaviour and so the traffic flow. On the one hand this is the car-following model which describes the acceleration and deceleration behaviour of the vehicles. And on the other hand a lane-change model which describes the way vehicles will change their lane due to the traffic situation. The reactions of these models are calculated upon the behaviour of surrounding vehicles.

3.1 Microscopic traffic simulation based on SUMO

The Open Source Software *SUMO* was founded in 2002 at the German Aerospace Center DLR [Ber11]. The main concept was to create a traffic micro simulation which is as fast as possible and provides easy opportunity of extension. The software is written in *C++* and uses some libraries to manage several tasks e.g. XML handling. Several projects with *SUMO* in the academic field have shown its suitability for academic purposes. To get an exact model of the real traffic scenario it is useful to integrate real-time traffic information from various sensors [Kru13c]. To calibrate the distribution of the vehicle fleet it is valuable to use the captured data of automatic traffic counters. Furthermore an approach to insert live traffic light data into the simulation is conceivable. The sensor data can either be used as an input source for the simulation e.g. traffic volume, traffic distribution or for validation purposes for example the journey times for specific routes.

3.2 Interaction with SUMO during the Simulation

For many research studies in the field of traffic simulations it appears to be useful to interact with a simulation at runtime. For that purpose the microscopic traffic simulation software *SUMO* has a real-time I/O data interface (*TraCI*) implemented. This interface offers the possibility for a bidirectional communication between the user application and the simulation. In the simulation suite a native Python Client application is available. Furthermore another application was designed to offer nearly all *TraCI* functions within a Web service. This software is named "*TraCI* as a Service – *TraaS*" [Kru13b] and it is published under the terms of the General Public License. The implementation of the Web service was done with the programming language *Java* which offers two different possibilities of usage [Kru13a]. *TraaS* can either be used as a stand-alone Web service with multiple clients (Figure 1) or as a native *Java* library.

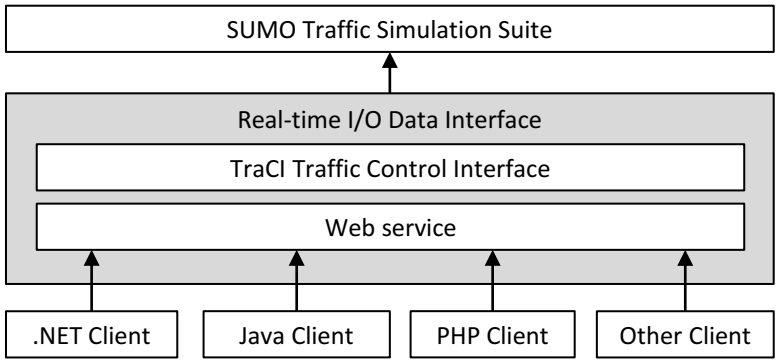


Figure 1: Structure of the communication framework.

For the integration in MATLAB the usage as library is convenient due to the fact that *MATLAB* is based on *Java* and offers good opportunities for *Java* Code integration. At the current state over 200 functions are available for getting or setting different parameters of the simulation.

3.3 Bidirectional Communication between SUMO and MATLAB

The *TraaS* Library has to be integrated into the *MATLAB* Workspace to use the *Java*-methods to interact with parts of the *SUMO* simulation. Beneath *MATLAB* a compatible *Java* Runtime Environment (JRE) is necessary. After importing the *TraaS*-Library a bidirectional communication between *SUMO* and *MATLAB* becomes available. As shown in Figure 2 the *TraaS* Library is embedded in *MATLAB* to access all the methods provided by *TraaS* (further information on methods provided by *TraaS* is available in the Documentation [Kru13b]). With this framework it is possible for *MATLAB* as a client to open a new *SUMO* instance as *TraCI*-server and establish a connection for communication.

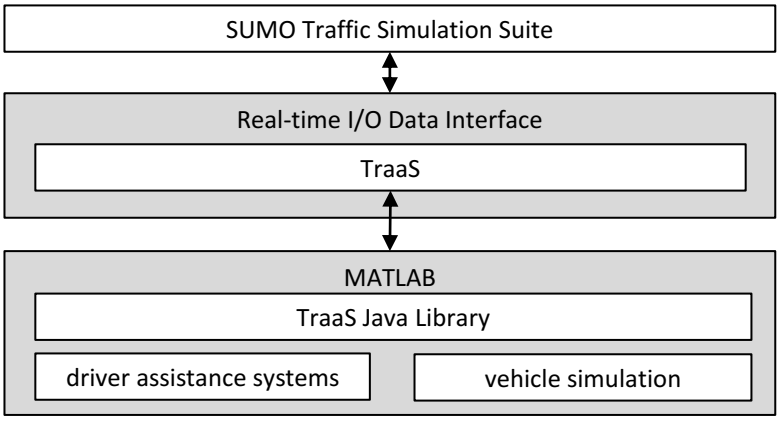


Figure 2: Structure of the communication framework between *SUMO* and *MATLAB*.

ADAS need some necessary inputs from *SUMO* to compute the speed profile of a vehicle. In Section 4 the simulation framework will be used to analyse a traffic light assistant system (TLAS) described in [Sch10]. Hence, an optimal speed profile for a vehicle driving towards a

traffic light is calculated on the basis of various inputs (e.g. current driving speed, maximum permitted speed, distance to next traffic light, time until next red or green phase, current queue length). At the current state *TraaS* does not provide this information. Consequently some advanced functions (see Table 1) are created in *MATLAB* based on the fundamental *TraaS Java* methods.

Table 1: Some advanced functions implemented in *MATLAB* using the *TraaS*-Library.

function	description
getLSAInfo()	returns time until the next green and red phases
getLSADistance()	returns the distance from vehicle position to next traffic light
getLSAOnRoute()	returns an list of all traffic lights on Route

With those new implemented functions it is possible to interact with certain *SUMO* specific simulation parts. An excerpt of the functionality is listed below:

- Initialising a new simulation instance including the *TraCI* Server,
- Managing the simulation process (configure, start, stop, pause simulation, etc.),
- Getting data from vehicles, lanes, edges, traffic light systems,
- Sending data to manipulate vehicles, lanes, edges, traffic lights, etc.
- Computing the distances to next traffic lights and
- Computing the duration until the next red and green phases.

These functions are constantly being extended with regard to the claims of current research projects. This framework allows a wide range of future analyses on advanced driver assistance systems and also on detailed vehicle simulation in the context of traffic scenarios.

4 First Application of the Framework

To illustrate the framework and to show the benefit of the simulation environment a TLAS [Sch10] is implemented in *MATLAB*. A road network with a single intersection controlled by a fixed traffic light program is generated. The incoming roads at the intersection consist of two lanes. The traffic flow approaching the intersection is modelled as a continuous stream without groups of vehicles. A parameter variation was done by simulating different traffic scenarios.

As already described in Section 3.3 the TLAS calculates an optimal speed on the basis of various inputs listed in Table 1. After simulating different scenarios the results can be analysed and visualised. In Figure 3 some vehicles (solid blue line) driving toward the traffic light which is located at position *s=1000 m* pass through the stretch without stopping at the intersection. At position *s=530 m* they get a speed advise by the TLAS and reduce their speed to the target value. The resulting speed profiles are used for the simulation in *Simulink*. The energy consumption of each driving situation is determined by using a model of an electric vehicle. A special component implemented in this model is the recovery of brake energy (recuperation).

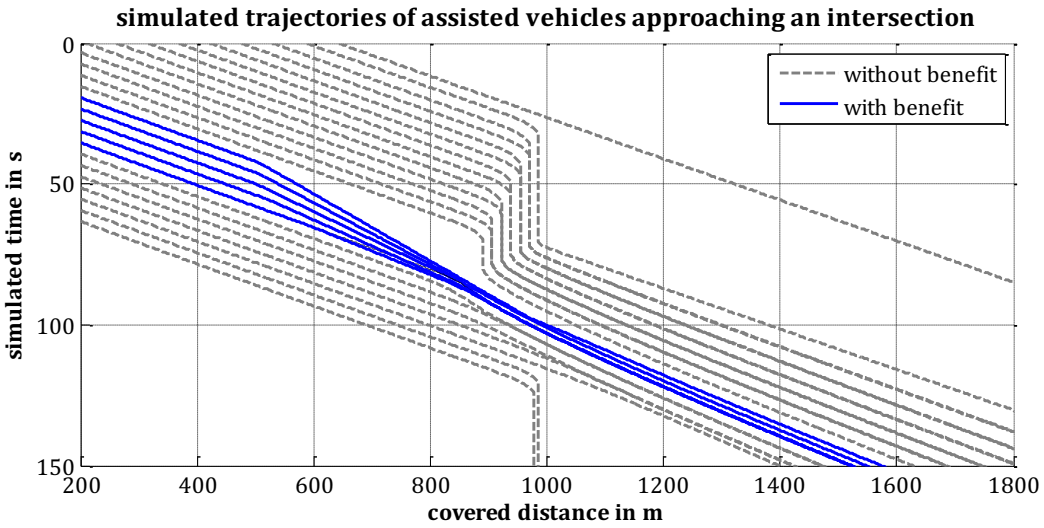


Figure 3: Various driving situations in front of a traffic light

In Figure 4 the difference in speed and energy consumption between assisted and not assisted situation is illustrated. It can be seen that the assisted driver reduces the speed at position *s=530 m* to approx. *v=35 km/h*. This permits a significant recuperation of brake energy. Also a lower energy consumption of the assisted vehicle until a covered distance of *s=900 m* is illustrated. This leads to a clear reduction in energy consumption for the described scenario. As a first result it is clearly visible that a TLAS has a significant benefit in numerous situations of vehicles driving toward a traffic light. Indications for the potential of energy recuperation while braking can be seen.

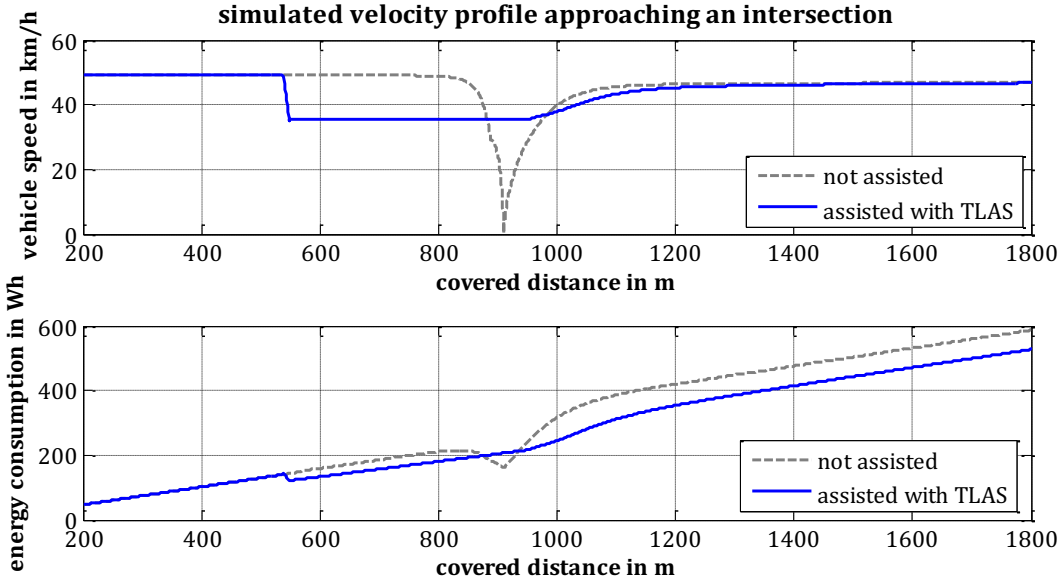


Figure 4: Example for vehicle speed calculated in *SUMO* using a TLAS programmed in *MATLAB*. Not assisted situation compared to assisted situation.

With the framework it is easy to change different parameters like speed limits, traffic light control, number of lanes, density of traffic flow and many more. Also the impact of different traffic situations on the energy consumption of the whole traffic can be determined. With the integration of *TraaS* into the Workspace of *MATLAB* a very powerful way for the analysis of traffic scenarios is created.

5 Conclusion

This contributed work is to be understood as an excerpt of the extensive possibilities that arise through the use of this new simulation environment. The framework between *MATLAB* and *SUMO is described*. It gives a first impression on how it works and what is intended to do with it. First results of the usability and the effect of a traffic light assistance system are shown. In further research studies the benefits will be evaluated with the described framework. However many aspects towards improved usability in further research projects are still in development.

Currently the energy consumption is determined by simulating vehicle models in post-processing. In the future it will be possible to simulate *Simulink* models during traffic simulation in *SUMO* to evaluate and optimise control strategies for e.g. hybrid electric vehicles or trucks mentioned in [Kut11] and [Tem12]. For simplification of the applicability and usability a *Simulink* library is built [Kot13] which allows users to integrate their models into the traffic flow simulation.

References

[Ber11] M. BEHRISCH, L. BIEKER, J. ERDMANN, and D. KRAJZEWICZ: "SUMO - Simulation of Urban MObility: An Overview". In: *SIMUL 2011, The Third International Conference on Advances in System Simulation*. 2011. (Last Access: 3 June 2013). URL: http://sumo.sf.net/pdf/simul_2011_3_40_50150.pdf.

[Bia05] M. BIAN: "A velocity control strategy for vehicular collision avoidance system". In: *2005 IEEE International Conference Mechatronics and Automation.* Vol. 4. July 39–Aug. 1, 2005, pp. 1827–1830.

[Dor04] C. DORRER: "Effizienzbestimmung von Fahrweisen und Fahrerassistenz zur Reduzierung des Kraftstoffverbrauchs unter Nutzung telematischer Informationen". In: *Schriftenreihe des Instituts für Verbrennungsmotoren und Kraftfahrwesen der Universität Stuttgart.* 24. Edition. Renningen, 2004. ISBN: 3-8169-2384-4

[FKA13] FORSCHUNGSGESELLSCHAFT KRAFTFAHRWESEN MBH AACHEN: *Produktlösungen – PELOPS: Systematische Beschreibung*. (Last Access: 18 June 2013). URL: http://www.pelops.de/

[Hab11] J. HABER-KUCHARSKY: *Traffic Simulation Toolbox - User's Manual*. Waterloo, Canada: University of Waterloo, Department of Electrical and Computer Engineering, May 31, 2011.

[INF13] INFRAS: *The Handbook of Emission Factors for Road Transport (HBEFA)*. (Last Access: 7 June 2013). URL: http://www.hbefa.net/e/index.html.

[IPG13] IPG AUTOMOTIVE GMBH/AVL: *IPG CarMaker for Hybrid and Fuel Consumption. CarMaker/Cruise - The integrated solution for hybrid and fuel consumption optimization*. (Last Access: 10 April 2013). URL: http://www.ipg-automotive.com/index.php?id=532.

[Kok11] Z. KOKKINOGENIS: "Towards the next-generation traffic simulation tools: a first evaluation". In: *DSIE'11 - 6th Doctoral Symposium on Informatics Engineering.* 2011.

[Kot13] S. KOTRBATY: "Entwurf und Realisierung einer Datenschnittstelle zur gekoppelten Simulation von SUMO und MATLAB/Simulink". Report, student research project. Dresden: Technische Universität Dresden, 2013.

[Kru13a] M. KRUMNOW: "Sumo as a Service – Building up a Web service to interact with SUMO". In: *Proceedings of the 1st SUMO User Conference SUMO2013, Reports of the DLR-Institute of Transportation Systems.* Vol. 21. Berlin, May 15–17, 2013.

[Kru13b] M. KRUMNOW: *TraaS – TraCI as a Service, An Extension to communicate with sumo based on the SOAP protocol*. (Last Access: 10 June 2013). URL: http://traas.sf.net/.

[Kru13c] M. KRUMNOW: "Microscopic real-time simulation of Dresden using data from the traffic management system VAMOS". In: *19th ITS World Congress.* Vienna, Austria, Oct 25, 2012.

[Kur13] T. KURCZVEIL and E. SCHNIEDER: "Implementation of an Energy Model and a Charging Infrastructure in SUMO". In: *Proceedings of the 1st SUMO User Conference SUMO2013*. Reports of the DLR-Institute of Transportation Systems. Vol. 21. Berlin, May 15-17, 2013, pp. 88–94.

[Kut11] S. KUTTER and B. BÄKER: "An iterative algorithm for the global optimal predictive control of hybrid electric vehicles". In: *7th IEEE Vehicle Power and Propulsion Conference (VPPC).* Chicago, 2011.

[Mac13] J. MACEDO, G. SOARES, Z. KOKKINOGENIS, D. PERROTTA, and R. ROSSETTI: "A Framework for Electric Bus Powertrain Simulation in Urban Mobility Settings: coupling SUMO with MATLAB(R)/Simulink nanoscopic model". In: *Proceedings of the 1st SUMO User Conference SUMO2013*. Reports of the DLR-Institute of Transportation Systems. Vol. 21. Berlin, May 15-17, 2013, pp. 96–102.

[Mai11] R. MAIA, M. SILVA, R. ARAÚJO, and U. NUNES: "Electric Vehicle Simulator for Energy Consumption Studies in Electric Mobility Systems". In: *Integrated and Sustainable Transportation System (FISTS), 2011 IEEE Forum*. Coimbra, Portugal, June 29–July 1, 2011, pp. 227–232.

[MAT13] MATHWORKS: *MATLAB(R) – The Language of Technical Computing – Overview*. (Last Access: 30 May 2013). URL: http://www.mathworks.co.uk/.

[Qua13] QUADSTONE PARAMICS LTD.: *Paramics Software Suit.* (Last Access: 28 September 2013). URL: http://www.paramics-online.com/.

[Sch10] T. SCHUBERT: "Entwurf und Evaluierung einer prädiktiven Fahrstrategie auf Basis von Ampel-Fahrzeug-Kommunikationsdaten". Diploma thesis. Dresden: Technische Universität Dresden, 2010.

[Sch11] P. SCHURICHT, O. MICHLER, and B. BÄKER: "Efficiency-increasing driver assistance at signalized intersections using predictive traffic state estimation". In: *14th International IEEE Conference on Intelligent Transportation Systems (ITSC).* 2011, pp. 347–352. DOI: 10.1109/ITSC.2011.6083111.

[Som11] C. SOMMER, R. GERMAN and F. DRESSLER: "Bidirectionally Coupled Network and Road Traffic Simulation for Improved IVC Analysis". In: *IEEE Transactions on Mobile Computing* 10.1 (Jan. 2011), pp. 3–15.

[Tem12] C. TEMPELHAHN and B. BÄKER: "Optimality-based Generation of Speed Trajectories for Parallel-Hybrid Commercial Vehicles". In: *9th Symposium HEV 2012*. Braunschweig, Germany, Feb. 14–15, 2012.

[Ved13] M. VEDENEV: *Traffic Simulation in MATLAB*. (Last Access: 28 September 2013). URL: http://simulations.narod.ru/.

Corresponding author: Dipl.-Ing. Torsten Schubert, Technische Universität Dresden, "Friedrich List" Faculty of Transport and Traffic Sciences, Chair of Vehicle Mechatronics, Dresden, Germany, phone: +49 351 463 39567, e-mail: torsten.schubert@tu-dresden.de

On-line Traffic Modelling In Assen: The Sensor City

Klaas Friso[1], Kobus Zantema[1], Edwin Mein[2]

[1] Goudappel Coffeng BV

[2] Omnitrans International BV

Abstract

In the Dutch municipality Assen, a large national R&D project on sensor networks is going on. The main goal is to improve traffic flow and environment in the city centre and surrounding region, by developing real-time intelligent traffic management systems making use of different sensing sources of real-time traffic data. For this purpose, a real-time traffic estimation and prediction model will be developed and deployed. A novel calibration approach is developed using a macroscopic model of a larger region. Different monitoring data sources consisting of both fixed-point and floating car data are used. This offers new opportunities, such as better estimation of origin-destination matrices and simulation in urban environments, which will be used to give better travel information towards travellers concerning the traffic situation in the Assen region. This paper presents the implementation of the on-line forecasting model structures and the first findings on the model system design.

Keywords: Sensor City Assen, On-line traffic model, sensor networks, short-term prediction, Omnitrans, StreamLine, Rolling Horizon

1 Introduction

Sensor City Mobility is a striking innovative mobility project with one main goal: to facilitate travelers with a personal travel advice so that they can easily choose the most comfortable and smart way of traveling. The project aims at a revolution in traffic and travel information services through smarter use of data from sensor technology. From January until November 2013, the Dutch City of Assen is a 'living lab' where a large-scale practical experiment is conducted. In this experiment hundreds of travellers try new in-car services and smartphone apps. This project is implemented by a consortium of companies and government.

In this paper we will present the development of an on-line traffic model which continuously monitors the current traffic status in Assen and makes a forecast of the expected traffic situation for half an hour and the next hour. The model will be updated every 5 minutes for the prediction of half an hour period and separately on another machine every

10 minutes to predict until one hour ahead. The reason for this splitting of calculations is that it is (at this time) not possible to do the prediction up to one hour ahead every 5 minutes. The model forecasts in terms of expected traffic flows, travel times and remaining road capacities will be used as input for the in-car services and smartphone apps. The on-line traffic model is build in Omnitrans, the transport planning application designed for integrated modelling of multi-modal transport systems. The dynamic traffic assignment model used in Omnitrans is called StreamLine [RAA10], an approach based on the cell transmission model (CTM), which is a derivative of the LWR model of Lighthill and Witham [LIG55] and Richards [RIC56].. This framework is capable of integrating aspects like route choice effects, traffic management, blocking back effects and departure time choices in a single software suite.

2 Comparison with other real-time traffic models

Later in this paper we will present the framework and results of the traffic model of Assen using the Omnitrans StreamLine model software. But first we like to discuss comparisons and differences with other real-time systems available. A comparison is made with Aimsun Online and PTV Optima. This is a basic comparison on what we found in literature.

2.1 Aimsun Online

The architecture of Aimsun Online [AIM13] is comparable with our framework. The most important difference in the short term forecasting is that in Aimsun Online a pattern recognition module is used to select the best suitable OD-matrix from an historical OD-matrix database, where we use a matrix calibration module. In our opinion it is quite a challenge to possess a historical database which includes all possible day-to-day variability and therefore it is more appropriate to use a matrix calibration procedure.

2.2 Optima

The idea of the OPTIMA architecture [GEN11] is comparable with our approach. Also a matrix calibration procedure is used to estimate the current OD-matrix using traffic counts. The dynamic network loading is similarly based on kinematic wave theory (macroscopic flows) apart from the fact that OPTIMA used turn splitting rates where our approach uses route flows. The disadvantage of using splitting rates in a dynamic model is that it isn't ensured that all trips in the OD-matrix will reach their real destination.

3 Process description

In [KLU11] a sensor network design was described to determine the best possible sensor locations. This design was based on the static transport model of the city of Assen in such a way that (1) it captures a maximum number of OD pairs and trips; (2) it is located at major decision points or road sections; and (3) it avoids too much overlapping with adjacent sensors.

Based on local knowledge of the network situation and their limitations at some locations this technically derived network design was somewhat adjusted and in 2011 a start was made with implementing this sensor network. At this moment (June 2013) about 90% of the sensors planned are implemented and the live data streams are started and being tested. The implementation process has taken much longer than was foreseen at the start of the project in 2010. The following reasons for the delay can be put forward for this:

- Tender procedures
- Different providers of traffic lights data
- Technical problems in the construction of pipes
- Construction of the sensor network had to be matched with municipal road works.

The reasons mentioned above meant that the implementation of the sensor network could not always be done in the most efficient way and therefore took longer as planned. The most important lesson that can be learned from above is to be aware of possible delays (for different reasons: legal, technical and practical) in implementing such a big sensor network. Therefore a tight planning schedule is needed including risk management.

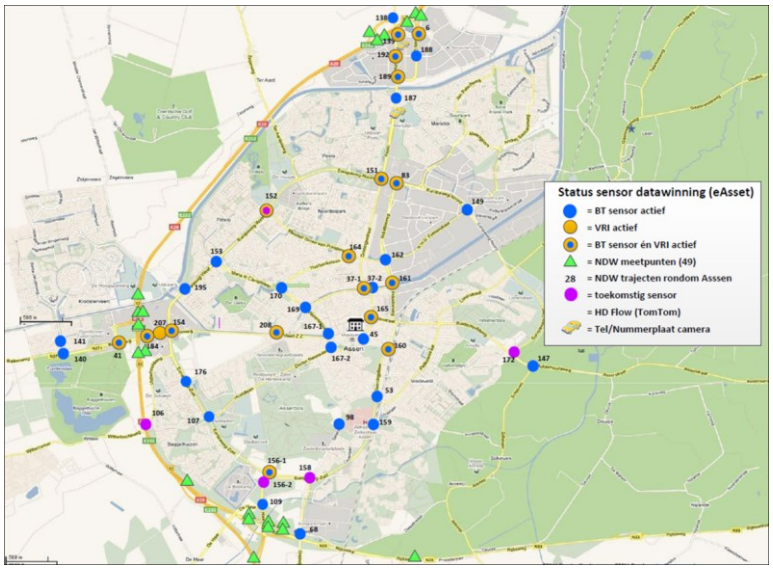


Figure 1: Sensor network Assen.

In Figure 1 an overview of the actual status of the sensor network in Assen is shown. The sensor network consists of 17 traffic lights, 6 cameras (2 for counting and 4 automatic number plate recognition) 49 highway loop detectors and 40 Bluetooth sensors. Furthermore, Floating Car Data will become available from navigation devices and OBU (On Board Units). From these devices different kind types of data become available like traffic flows, speeds, travel times and OD pairs. The on-line traffic model uses the traffic flows as input and the other types of data as validation tools. Important aspect in the data handling is the filtering of unreliable measurements. For example, in this urban area where with their mobile phones are also on the road or parallel cycling paths and will also be measured by Bluetooth sensors. These measurements have to be filtered out of the dataset.

4 Architecture on-line traffic model

The on-line traffic model is run every 5 minutes to make a traffic state prediction of up to 30 minutes ahead. On a separate machine a second run is made every 10 minutes to predict up to an hour ahead. From the traffic model flows, spare capacities and travel times are deduced. The architecture of the on-line traffic model consists of three basic components, which are run subsequently to make the state estimation. The overview of the architecture is displayed in Figure 2 below and described in detail thereafter.

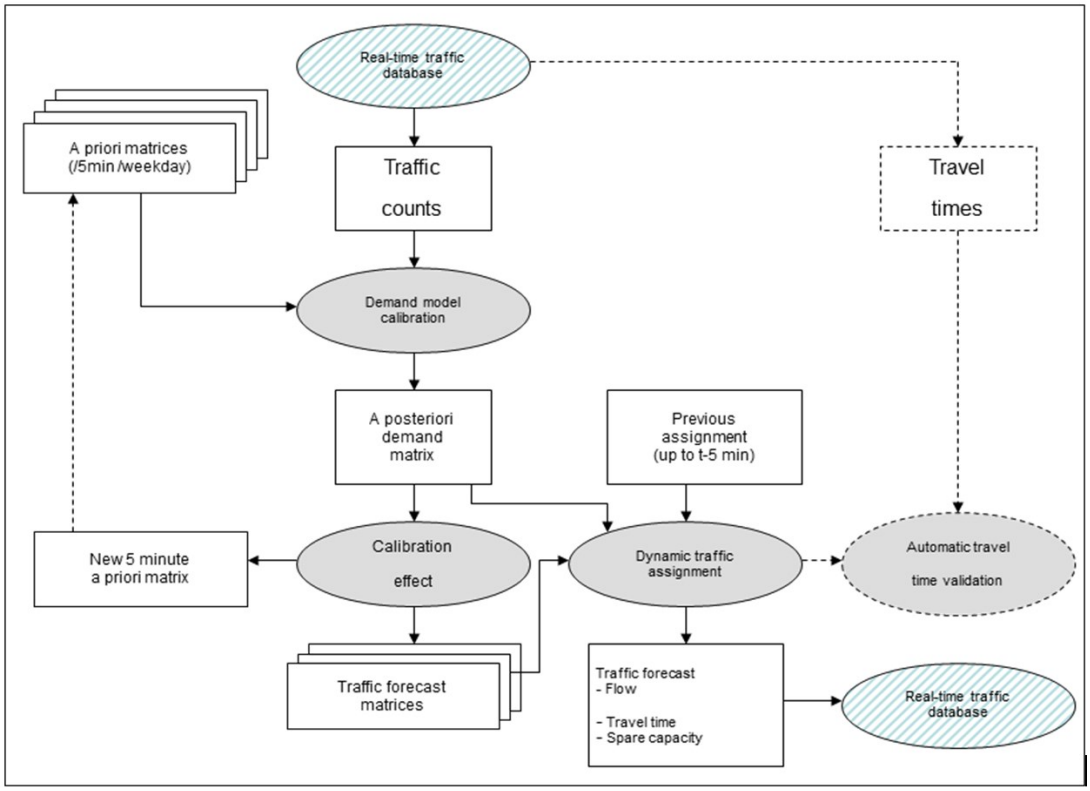


Figure 2: Architecture on-line traffic model.

At first the current traffic counts and travel times are taken from the Real-time traffic database within the Sensor City Assen project. These traffic counts contain loop data on the motorway, as well as loop data at traffic lights. In the future counting camera's and Bluetooth estimation may be added to this. The travel times are made up from Bluetooth measurements and TomTom floating car data. At the moment only the traffic counts are used in the online model calibration and forecasting.

After the traffic counts have been collected the current OD-matrices are picked up from a large collection of matrices. Within the project a matrix has been estimated for every 5 minutes of the week, resulting in a total of 2.016 base matrices. These base matrices have been estimated by adding a demand pattern over the original static matrices. The demand patterns are determined from yearly household travel surveys by OViN (Onderzoek Verplaatsingen in Nederland) [CBS13]. For the peak periods a different static matrix has been used than for the off-peak period, considering the different flow pattern between periods.

The next step is to calibrate the current demand matrix on the basis of the current traffic counts. The calibration procedure is a information minimization algorithm [ZUY80]. On the result the calibration effect is calculated and partially taken along on the future matrices to create the traffic forecast matrices. The calibration effect is in fact the difference between the actual situation and the average situation. This relative effect is also progressed to the short term forecast. At the moment a large part of the calibration effect is taken along in the forecast matrices. Starting with 95% for the first 5 minutes and decreasing continuously to 70% for 30 minutes ahead. The a priori matrices are continuously updated every 5 minutes by adding the calibration effect partly (initial value is 3%, which has to be evaluated further on) to the current a priori matrix, to let the system learn by itself.

With the current and traffic forecast matrices ready, the dynamic traffic assignment is executed, resulting on all links in traffic flows, speed, travel time and spare capacity. In order to keep the start-up phase of the dynamic assignment short, a rolling horizon method is used. Seeing the significance of this, it is described in more detail in section 4.1. After the dynamic simulation the results are send back to the real-time traffic database of Sensor City Assen, so that other applications can use this additional information, for example for Smart Routing purposes.

In the near future it is planned to include also an auto-validation of the calculated travel times on the basis of measured travel times. This is future research however, though it is likely to have been concluded before the conference in December 2013.

4.1 Rolling horizon

Predicting the traffic's condition requires a realistically, dynamically loaded network. Normally a DTA simulation starts with an empty road network and traffic starts flowing from all zones into the network, based on the OD-matrices, routes, route choices etc. In order to obtain a realistic traffic situation for a prediction, the network has to be loaded entirely, otherwise an unrealistic prediction would be made. In order to obtain a loaded network the DTA-simulation uses a pre-loading period prior to the actual simulation period.

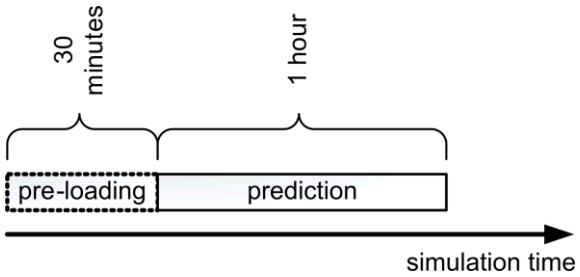


Figure 3: Pre-loading simulation.

The pre-loading depends on the network's size. In our case a 30 minute pre-loading simulation is required to obtain a fully loaded network prior to making the actual prediction of one hour. With an update interval of 5 minutes, this means that the same approach will be repeated every 5 minutes based on new measurements and OD matrices. This process has to be finished within 5 minutes in order to be ready for the next interval.

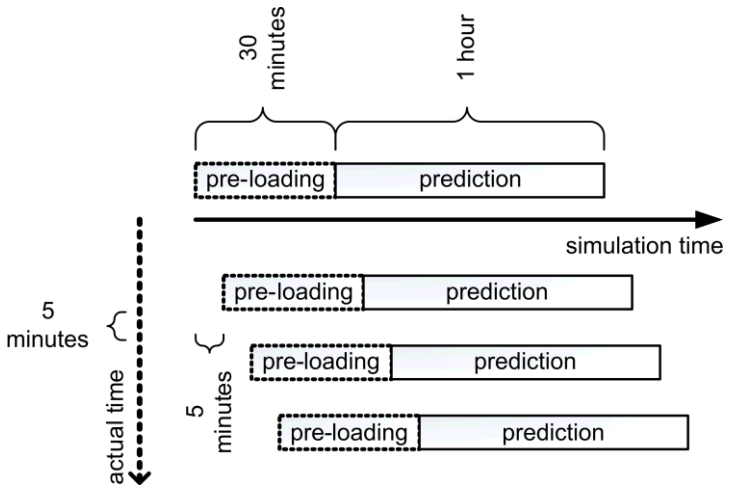


Figure 4: Simulation process.

It needs no explanation that the approach above using pre-loading consumes considerably amounts of computing time just for obtaining a realistic prediction every 5 minutes.

Observing the approach, one can argue that restarting the entire simulation including the pre-loading phase is not essential as long as it is possible to change the OD-matrices, routes and route-choice between each 5 minute-interval.

Our solution for this is the ability to store the simulation's condition at any time during the simulation. Therefore we have adapted StreamLine so it writes its internal situation to file. This internal situation represent everything the simulation knows about the condition on all links, link's cells, speeds, spill back, junctions, zones etc. This process is called serialization [Cog05] based on the Boost library.

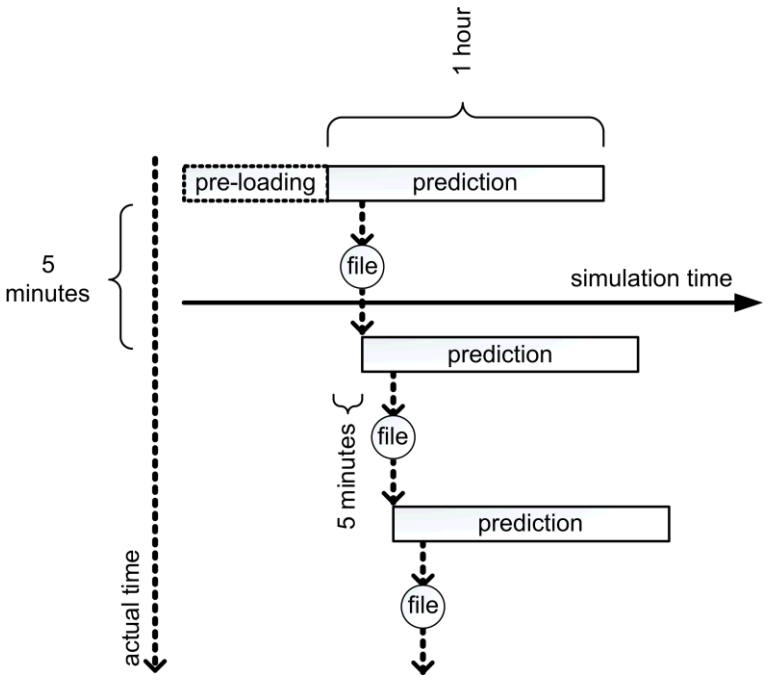


Figure 5: Serialization to save computation time.

Using the stored internal situation in a file on disk, StreamLine can reproduce that particular situation simply by reading the file and configuring all objects internally as if the

simulation never stopped. From that situation new OD-matrices, route choices etc. can be applied to the new simulation run. The process of using serialization eliminates the pre-loading phase and starts off with a realistically, dynamically loaded network. As can be seen, this saves enormous amount of computing time and enables StreamLine to run within the 5 minute update interval as discussed in this paper.

5 Results off-line model simulation

For the results of the model we will refer to a 1-day simulation based on real traffic data. In other words, as if the data is coming in online and used for the current demand model calibration and prediction of flows, travel times etcetera. Results are mainly presented as the effect of the use of data for the traffic forecast and the variation in the forecast compared to the current situation.

On a first positive note, with a near empty log file, the simulation was able to run well within 5 minutes for each time period. Since this will be essential during the online simulation it is just as essential to keep the log file small during simulation, or store it at a different location for each time period.

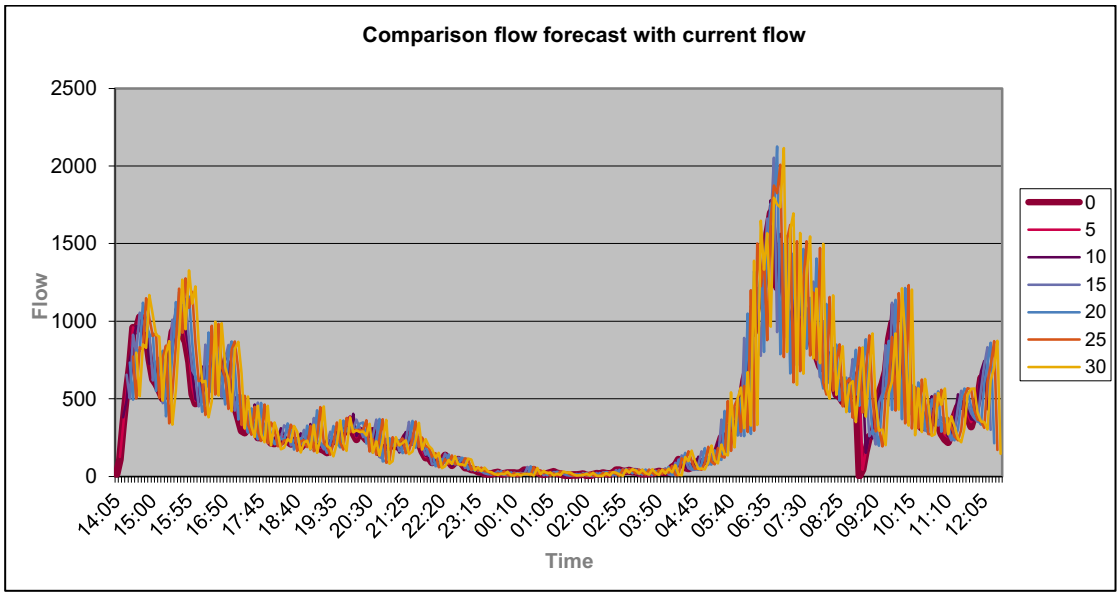


Figure 6: Result off-line simulation 22 hour period (Motorway A28).

Figure 6 and Figure 7 show results of the 22 hour simulation. In Figure 7 a more detailed picture is given, with the focus on the morning peak. Input for the graphs is a single motorway A28 link to the southwest of Assen. The fat line (0) is the flow at the current time, where the other 6 lines show the flow at the traffic forecast times, up to 30 minutes ahead.

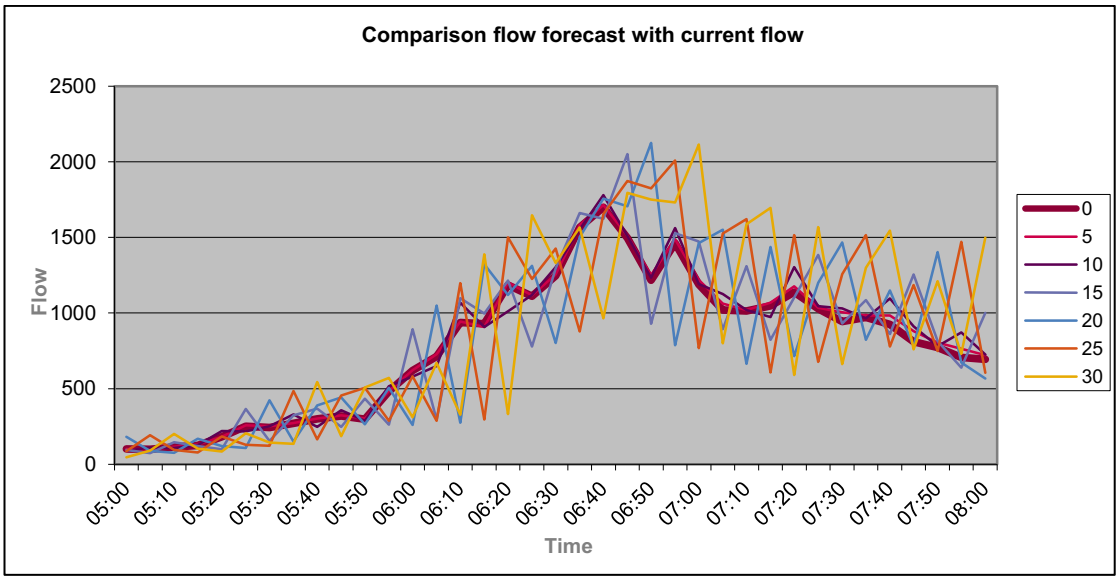


Figure 7: Result off-line simulation morning peak period (Motorway A28).

On a first impression the traffic forecast shows a lot of fluctuation. But also the current traffic isn't a 'smooth' line, something which is the case in the current base matrices. Since a large part of the calibration effect is currently taken into account, a slight upsurge from the traffic flow will result in the same upsurge in the forecast matrices. The same effect holds true if the flow is lower then would be estimated from the base matrices. Also, if there is a shift in the traffic demand a few minutes late compared to the reality, the flow for the upcoming half hour will be underestimated. As is clearly visible in Figure 7 around 6 am.

On a local road the same effects as for the motorway are found. In Figure 8 we once again focus on the morning peak period. As clearly shown a sudden rise or fall in the current flow has a strong effect on the traffic forecast, resulting in flow forecasts of nearly three times the volume found as the real time arrived (at 6:15 am) or more than half as low (at 7:15 am).

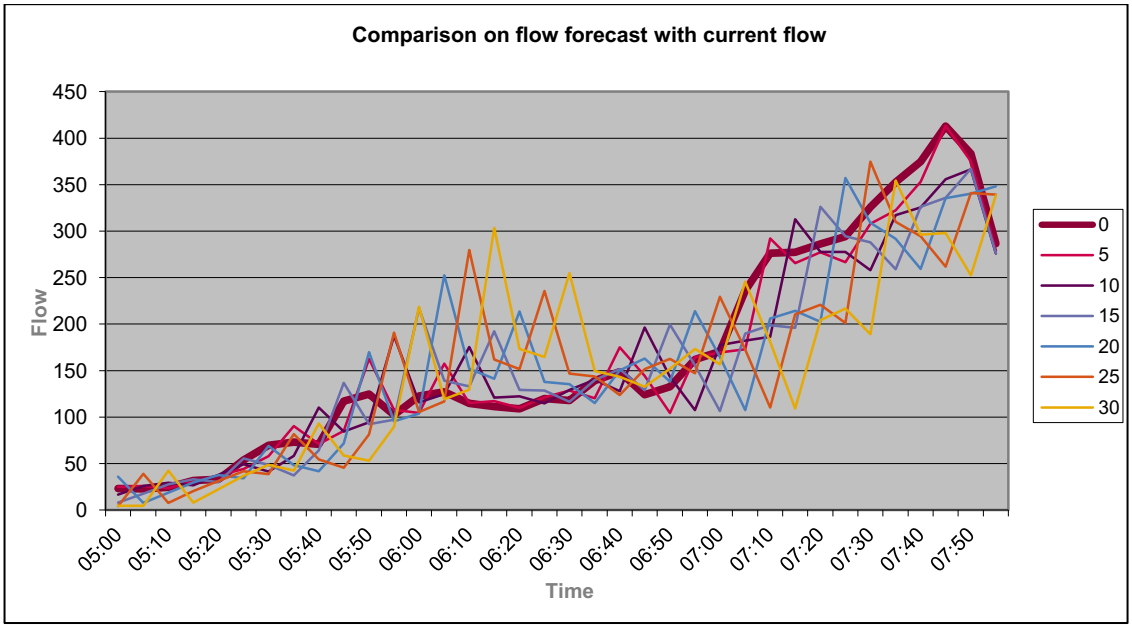


Figure 8: Result off-line simulation morning peak period (local road).

Clearly the pattern from the demand matrices does not yet match the pattern shown on the road. Taking into account the calibration effect partly in the base matrices will result in a better traffic pattern over time in a few weeks of online simulation and calibration, as the model system needs some time to learn the regular traffic patterns in Assen. Think of the fact that the current base OD-matrices are based on a static transport model that was build a couple of years ago and the demand patterns are based on household surveys, so it is therefore not surprising that a running-in period is needed to stabilize. Also it is shown that at current, the calibration effect has too much effect on the traffic forecast, and different calibration effect percentages as well as the muting (from 95% to 70%) as described in chapter 4 will be tested in the near future.

6 Conclusions

In this paper the first results from an offline simulation in an online model environment are presented. The model is clearly technical operational and the first results are promising. The upcoming months the online model will go 'live'. Additional research is needed on calibration of the model parameters. In particular the optimal percentage calibration effect applied on the short term prediction matrices and updating the set of a priori matrices ('the learning principles') will be determined. Also a monitoring module will be implemented where besides the comparison of predicted flows with observed flows also the predicted speeds and travel times will be compared with observed speeds and travel times.

Acknowledgement

We acknowledge the Stichting Sensor City for giving us the opportunity to work on this project and the city of Assen of providing us with real-time data. The Sensor City project is being facilitated by the European Union, the European Regional Development Fund, the Ministry of Economic Affairs, Agriculture & Innovation and The Northern Netherlands Provinces, KOERS NOORD. Furthermore we like to thank the anonymous reviewers for their useful comments.

References

[Aim13] *Aimsun online*. 2013. URL: http://www.aimsun.com/wp/?page_id=33.

[CBS13] *OVin, Onderzoek Verplaatsingen in Nederland*. 2013. URL: http://www.cbs.nl/nl-NL/menu/informatie/deelnemers-enquetes/personen-huishoudens/ovin/doel/default.htm.

[Cog05] J. Cogswell: "Adding an Easy File Save and File Load Mechanism to Your C++ Program". In: *InformIT*. July 1, 2005. URL: www.informit.com/articles/article.aspx?p=398702.

[Gen11] G. Gentile and L. Meschini: "Using dynamic assignment models for real-time traffic forecast on large urban networks". In: *MT-ITS*. Leuven, Jun. 22–24, 2011.

[Klu11] G. KLUNDER, Y. S. CHEN, K. ZANTEMA, and K. FRISO: "Consistent traffic Modeling and calibration at different resolution levels based on various sensor data for real-time traffic management". In: *MT-ITS*. Leuven, Jun. 22–24, 2011.

[Lig55] M. LIGHTHILL and G. WITHAM: "On Kinematic waves II: A theory of traffic flow on long crowded roads". In: *Proceedings of the royal society of London.* Part A 229.1178 (1955), pp. 281–345.

[Raa10] M. P. H. RAADSEN, H. E. MEIN, M. P. SCHILPZAND, and F. BRANDT: "Implementation of a single dynamic traffic assignment model on mixed urban and highway transport networks including junction modelling". In: *Proceedings of the third international symposium on dynamic traffic assignment.* Takayama, Japan, 2010.

[Ric56] P. RICHARDS: "Shock waves on the highway". In: *Operations Research* 4 (1956), pp. 42–51.

[Zuy80] H. J. VAN ZUYLEN and L.G. WILLUMSEN: "The most likely trip matrix estimated from traffic counts". In: *Transportation Research Part B: Methodological* 14.3 (1980), pp. 281–293.

Corresponding author: Klaas Friso, Goudappel Coffeng, Deventer, the Netherlands, phone: +31 570 666812, e-mail: kfriso@goudappel.nl

Reducing the Impact of Traffic Incidents Using Capacity-Regulating Traffic Lights

Markus Rausch, Stefan Lämmer, Martin Treiber
Technische Universität Dresden

Abstract

Traffic incidents are spontaneous events that obstruct the vehicle flow in road networks and might lead to significant congestion. In order to complete the set of countermeasures, which typically include the information of drivers, we propose to additionally regulate intersection capacities by the use of traffic lights in such a way that an incident-regarding traffic flow distribution in the network is attained. Provided that the incident is identified, the capacity-regulating traffic lights give shorter green times to flows towards congestion, while they expand the green times for those flows that bypass congested areas. The effective green times are obtained by an iterative optimization of signal plans with respect to the disturbed network and the anticipated travel times therein. We demonstrate the proposed method numerically for two distinct incident scenarios in a simple road network and also in the main road network of Dresden (Germany). In both networks, we reduce the impact of incidents in terms of travel time increase significantly as compared to the unregulated network.

Keywords: traffic incidents - traffic light controls - traffic management - gridlock prevention

1 Introduction

In recent times, the motorized individual transport increasingly gained influence on people's lives. However, the increased traffic demand often stresses even fully developed infrastructures and causes drivers to spend more time on reaching their destinations than would be necessary in light traffic [Sch12]. In addition, spontaneous events that affect the free flows of traffic, called *incidents*, induce congestion and further decrease the performance of those infrastructures. The National Traffic Incident Management Coalition estimates that incidents account for roughly $25\,\%$ of the overall congestion on U.S. roadways, suggesting to put more efforts into the management of those incidents.

Following the Traffic Incident Management Handbook [Uni00], we define incidents as non-recurring events that, on short time scales, cause either a reduction of the road capacity (e. g. lane blocking) or an increasing demand (e. g. mass events). If the demand exceeds

the capacity, incidents result in congestion that typically spreads over larger parts of the network, even although unused capacities are still available elsewhere [Dud75]. Ultimately, incidents can lead to traffic breakdowns or, yet worse, to gridlock situations in which the accumulation of cars in the network restricts potential outflow capacities [Dag07]. Since traffic flows are obstructed and infrastructure is less accessible, congestion dissolves only slowly after its cause (e. g. the incident) is no longer active. While the incident itself is not directly controllable, its impact can be reduced by applying suitable countermeasures that inhibit as well as relieve congestion. In particular, we consider driver information combined with traffic management.[1]

Being a precondition for the application of countermeasures, much efforts has been put into methods to detect incidents. Trivedi, Mikic and Kogut [Tri00], for instance, proposed to use distributed video camera networks. Concurrently, a number of authors developed algorithms based on detector data, e. g. Samant and Adeli [Sam00], Tang and Gao [Tan05] and Gall and Hall [Gal89], or based on floating car data, e. g. Kerner et al. [Ker05]. Detection methods cannot actively influence the transport scene but allow, amongst others, the application of driver information. Abuelela, Olariu and Weigle [Abu08], for example, proposed a notification system in which cars bidirectionally communicate with sensor belts installed on the road, allowing to detect incidents and, if actually detected, to disseminate information about their occurrence among the drivers. In this way, drivers are enabled to decide for bypass routes and to escape potential jams, thereby reducing the impact of this incident. This concurs with the investigations of Arnott, de Palma and Lindsey [Arn91] and Wunderlich [Wun98] who found that driver information in itself may lead to lower travel times.

We go one step further and propose an active traffic management during incidents by changing the signaling of existing traffic lights on an event-critical basis. Assuming that drivers tend to adopt a new user equilibrium (supported by driver information), we expect a significant reduction of the average travel time.

The present paper is organized as follows: In the first section, we elaborate on our proposition to utilize traffic lights regulating capacities during incidents. Further, we present our method to adjust the signal plans in disturbed networks, followed by simulation results. Subsequent to a discussion, future prospectives are provided.

2 Capacity Regulation During Incidents

We follow the idea that the impact of incidents could be reduced if green times of intersections in or near affected areas are reallocated such that more capacity is given in directions circumventing the congested area. Drivers can then utilize available bypass routes and, thus, experience a reduction of the travel time increase. Simultaneously, longer red times for flows heading towards the incident relieves the area from further inflows.

[1] Concurrently, the Traffic Incident Management Handbook (2000) [Uni00] has a more general perspective and summarizes (1) Detection, (2) Verification, (3) Driver information, (4) Response, (5) Site management, (6) Traffic management and (7) Clearance as incident management activities.

Generally, incidents induce a process in which the distribution of traffic flows drifts away from a present user equilibrium state towards a new state in which traffic flows utilize remaining road capacities on bypass routes. This transition is supported by informing the drivers about the incident. However, if signal plans do not adequately serve bypassing traffic, the efforts put into informing the drivers are largely given away. For that reason, we propose to simultaneously apply adapted signal plans right after the incident has occurred. This corresponds to a purposeful provision of capacity on bypass routes that are beneficial to be used.

As the essential idea is to regulate the capacity of intersections during incidents, we do not consider the correlation of signal timings between intersections (such as green waves) for which the effect is limited in disturbed networks anyway. Consequently, the regulation of capacities is completely realized by changing the green time fractions (splits) at affected intersections while keeping their cycle times fixed. We specify signal plans by minimizing the average travel time concerning all OD pairs in the network under the present incident. To this end, we use an iterative method illustrated in the next section.

3 Signal Timing Adjustment

The evaluation of adequate signal plans can be accomplished by minimizing the average travel time T on the condition of a present user equilibrium. The assumption of a user equilibrium can be justified by considering informed drivers. We employ the following objective function

$$T = \frac{\sum_j Q_j T_j}{\sum_j Q_j}, \tag{1}$$

which concerns the average travel time T_j and the demand Q_j of all OD pairs j. According to Wardrop [War52], all used routes r of OD pair j have the same average travel time T_r such that

$$T_j = T_r = \sum_{i \in r} T_i + \sum_{n \in r} T_n.$$

Consequently, T_r is the sum of the average travel times needed to pass all road segments i, T_i, and intersections (node) n, T_n, that are part of the route r, respectively. While the route fractions ω_{rj} largely influence T_i, the green times $\tau_{g,p}$ of phase p largely influence T_n. For modeling the average travel time T_i on road segment i, we employ a capacity-restraint (CR) function that was developed by the Bureau of Public Roads (BPR) and is given as follows:

$$T_i = T_i^0 \left(1 + \left(\sum_j \sum_{r(j)} \delta_{i,r}\, \omega_{rj} icl \frac{Q_j}{C_i} \right)^{\gamma} \right),$$

where T_i^0 is the free travel time on link i, $r(j)$ is the set of routes related to OD pair j, $C_i = Q_i^{\max}$ is the road capacity, and $\delta_{i,r}$ is 1 if route r proceeds via road segment i, and zero otherwise. In the present study, we use the parameter value $\gamma = 2$. On nodes, CR functions

Table 1: Reference signal plan of the undisturbed simple network defined by the fraction of the green times $\tau_{\mathrm{g}}/\tau_{\mathrm{c}}$ per intersection (IS).

IS	$\tau_{\mathrm{g}}/\tau_{\mathrm{c}}$ (WE)	$\tau_{\mathrm{g}}/\tau_{\mathrm{c}}$ (NS)
I1	21.67 %	78.33 %
I2	78.33 %	21.67 %
I3	50.00 %	50.00 %
I4	78.33 %	21.67 %
I5	21.67 %	78.33 %

are applied to every turning direction where values of T_n are estimated with the aid of an interpolation method according to the Intersection Capacity Analysis (ICA) [Boa10].

We perform our calculations by using the commercial transportation planning software PTV Visum (version 12.0). We assign fixed steady-state traffic demands Q_j to our road network and obtain average travel times T_j for all OD pairs j of the regarded network. Starting with the present signal plans (*reference*) that were applied before the incident has occurred, we obtain the *adapted* signal plans by iteratively varying the green-time fractions until the objective function in Eq. (1) is minimal with respect to the user equilibrium assignment. While we have reassigned the traffic demand to the network with Visum, one can also consider to use offline-optimization tools such as *TRANSYT* [Rob69]. Note that the adapted signal plans are only optimized for the regarded incident and might no longer be optimal after the incident became inactive. By comparing the average travel times T in the disturbed network, based on either the reference or the adapted signal plans, respectively, one should find that

$$T_{\mathrm{adapted}} < T_{\mathrm{reference}}$$

using Eq. (1). By finding this inequality fulfilled, the application of adapted signal plans can reduce the incident-related increase of travel times.

4 Simulation Results

We considered two example networks (Fig. 1), a simple artificial urban road network and the main road network of Dresden (Germany). While the simple network is considered in detail, we only give a summary of the results for the Dresden network.

4.1 The Simple Network

12 non-trivial OD pairs are connecting the source/sink districts 1–4. For simplicity, we assume both a symmetrical network where all links have the same capacity of $C = 1800$ vehicles per hour, and a symmetrical demand structure by assigning a stationary demand $Q_j = 333\,{}^{\mathrm{Veh}}/_{\mathrm{h}}$ to each OD pair. All five intersections I1-I5 are equipped with fixed-time traffic light controls

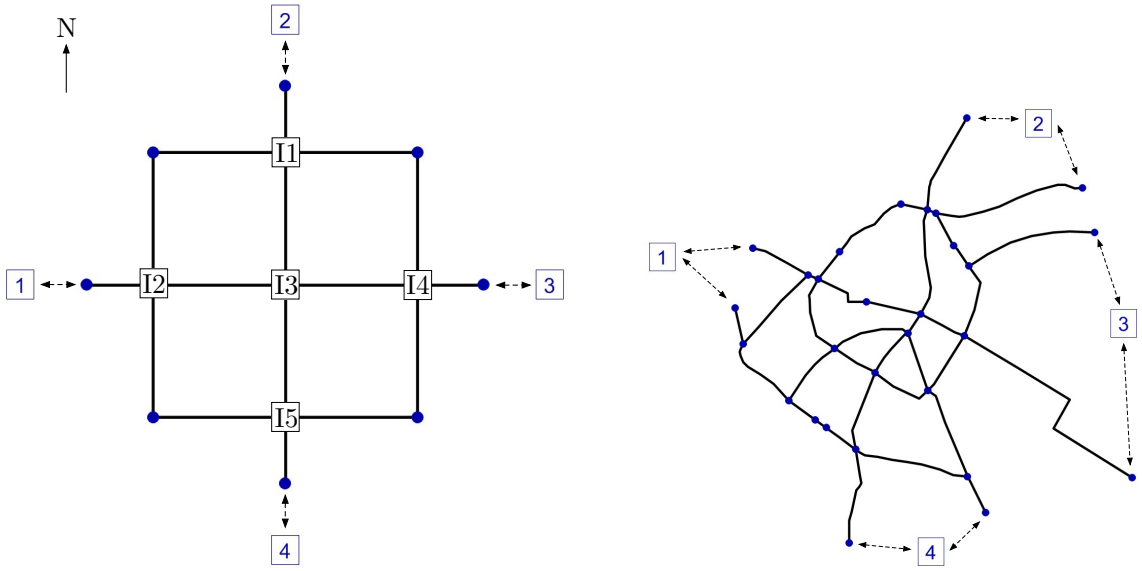


Figure 1: Urban road networks we considered in our study. For the left network, we explicitly show the iterative signal plan optimization for our two regarded incident scenarios. The improvements are principally the same in the more complex network of Dresden (Germany) on the right side.

with a common cycle time of $\tau_c = 60\,\mathrm{s}$, neglecting intergreen times and phase offsets. Every intersection has four main flow directions controlled by two phases: Phase WE for the west-east axis and phase NS for the north-south axis, respectively. Turnings to the left and right are generally permitted but not explicitly controlled; nor exist exclusive turning lanes. The road length between two adjacent nodes is $1\,\mathrm{km}$.

For reasons of symmetry, the optimized fractions of green times for the undisturbed network (and also for Incident A) belong to the following one-parameter family:

$$\begin{aligned}
\frac{\tau_{g1}}{\tau_c}(\mathrm{WE}) &= \frac{\tau_{g5}}{\tau_c}(\mathrm{WE}) = \alpha, \\
\frac{\tau_{g2}}{\tau_c}(\mathrm{WE}) &= \frac{\tau_{g4}}{\tau_c}(\mathrm{WE}) = 1 - \alpha \\
\frac{\tau_{g3}}{\tau_c}(\mathrm{WE}) &= \frac{1}{2}.
\end{aligned}$$

Evidently, the green fractions of the WE and NS phases add to unity.

The Undisturbed Simple Network

To set the reference to which the adapted signal plans are compared to in the disturbed network, we evaluated optimal signal plans in the undisturbed network as displayed in Table 1. With these reference signal plans, we obtain an average travel time of $T_{\mathrm{ref}}^{\mathrm{undisturbed}} = 6.0\,\mathrm{min}$ in the undisturbed network. We refer to any value that a driver has to spend longer on reaching his or her destination as travel time increase (TTI).

Table 2: Optimized green splits τ_g/τ_c per intersection (IS) in case of incident A (left) and incident B (right) in the simple network.

IS	τ_g/τ_c (WE)	τ_g/τ_c (NS)
I1	60 %	40 %
I2	40 %	60 %
I3	blocked	
I4	40 %	60 %
I5	60 %	40 %

IS	τ_g/τ_c (WE)	τ_g/τ_c (NS)
I1	23.33 %	76.67 %
I2	96.67 %	3.33 %
I3	53.33 %	46.67 %
I4	76.67 %	23.33 %
I5	60.00 %	40.00 %

Incident A – Intersection Blockage

In the first incident scenario, we assume that the central intersection I3 is completely blocked. Thus, traffic cannot use the inner links but has to utilize the outer ones. Therefore, we aim for providing capacity to all routes proceeding along the periphery. Apparently, all remaining intersections (I1, I2, I4, I5) are affected by the incident. After readjusting the green time fractions, we obtain signal plans as summarized in Table 2 (left).

By applying the reference signal plans (Table 1), drivers experience an average travel time of $T^{\mathrm{A}}_{\mathrm{ref}} = 22.6\,\mathrm{min}$ which means a travel time increase of about 16 minutes. The adapted signal plans, that fully regard the incident, result in $T^{\mathrm{A}}_{\mathrm{adptd}} = 14.6\,\mathrm{min}$ and, thus, in a travel time increase of 8 minutes. Consequently, we could reduce the travel time increase by 8 minutes.

Incident B – Higher Demand between Two Districts

In the second incident scenario, the traffic demand between the districts 1 and 4 is increased by a factor of six. As there are no blockages, the infrastructure remains fully accessible. We firstly expected that only intersections I2, I3, and I5 are affected. We found, however, that intersections I1 and I4 indeed exert some influence on the average travel time. Thus, all intersection are considered for readjustment.

In order to serve the higher demand, we give significantly more green time (96,67 %) to the WE phase of intersection I2. Intersections I3 and I5 also obtain more green times for their WE phases. Additionally, we marginally increase the green times for the WE and NS phases of intersections I1 and I4, respectively. The optimal signal plans are shown in Table 2 (right).

The average travel time becomes $T^{\mathrm{B}}_{\mathrm{ref}} = 19.2\,\mathrm{min}$ using the reference signal plans. With the adapted signal plans however, drivers only experience an averaged travel time of $T^{\mathrm{B}}_{\mathrm{adptd}} = 14.3\,\mathrm{min}$. Taken together, we could lower the travel time increase by 4.9 minutes.

Table 3 summarizes the obtained values in both incident scenarios using the reference and adapted signal plans. The reduction of the travel time increase related to the travel time increase for the reference case yields a relative improvement of 48.2 % in incident scenario A and 37.1 % in incident scenario B, respectively.

Table 3: Comparison between the average travel times in the simple network for the incident scenarios A and B. The travel time increase (TTI) compared to the undisturbed network could be significantly reduced if incident-regarding signal plans were applied.

Travel Times	No Incident	Incident A	Incident B
Reference T_{ref}	6.0 min	22.6 min	19.2 min
Adapted T_{adptd}	–	14.6 min	14.3 min
Reduction of TTI	–	8.0 min	4.9 min

4.2 The Simplified Road Network of Dresden

The main road network of Dresden (Fig. 1, right) was modeled by fourteen signalized intersections and a total road length of approximately $40\ \mathrm{km}$. We chose a symmetrical demand structure containing four districts. Similarly, we simulated a blocked intersection and a sufficiently higher demand between two districts and iteratively optimized the signal plans of affected intersections. We found that the travel time increases could be reduced by approximately $50\ \%$ in both incident scenarios.

5 Discussion

Based on two specific incident scenarios, we anticipated an incident-regarding, user equilibrium traffic flow distribution and adapted the original signal plans of affected intersections such that the resulting traffic flow distribution is minimized in terms of the average travel time of all OD pairs. The reference signal plans, which were optimized for the undisturbed case, do not regard the incident and, if applied during the incident, give away capacities that are urgently needed for bypass routes even if drivers are fully informed and act accordingly. In comparison, the adapted signal plans fully regard the incident scenario and perform much better in terms of the average travel time (see Table 3). Therefore, we conclude that travel time increases due to incidents can be reduced by an event-driven traffic management. Moreover the vicinity of the incident is relieved from traffic if bypass routes are utilized. This prevents both further congestion and gridlock situations.

We have demonstrated this by examining two distinct incident scenarios in a simple artificial road network and obtained reductions of travel time increase by approximately $40\ \%$ and $50\ \%$, respectively. However, in order for the measures to be effective, the traffic management must be accompanied by driver information to accelerate the shift towards the new traffic flow state. Moreover, we applied similar incident scenarios to the main road network of Dresden. Here, we observed that the travel time increases reduced by $50\ \%$. We therefore conclude that, if drivers are sufficiently informed, green time reallocation during incidents also performs very well in more complex networks.

In this paper, we demonstrated the feasibility of lowering the negative consequences of

incident-related congestion by adapting the signal plans of affected intersections. Although the state of information among drivers may vary significantly in reality, we assumed a present user equilibrium for our simulations. The impact of the idea presented in this paper strongly relies on driver information as it accelerates the redistribution process in case of an incident. Thus, although the traffic flow distribution is far from being in a user equilibrium in incident situations, informed drivers are able to quickly respond to it such that a new user equilibrium can be established.

Notice that our presented results serve as a best case estimate. They are only valid after the user equilibrium traffic flow distribution is fully reattained. Quantitatively more accurate results could be obtained by using a dynamic assignment procedure that also covers congestion dynamics.

A different aspect concerns network robustness. In simple terms, network robustness measures the ability of a network to cope with exceptional conditions like incidents. Snelder [Sne10] formulated, besides other measures, that a fastest-possible re-routing of drivers to alternative routes (if available) makes the network more robust against the incident. By informing the drivers and adapting the signaling of affected intersections, our proposed incident management pursues this objective. However, the improvement of network robustness by driver information crucially depends on the driver's response to the information [Li08].

It remains to be shown if the optimization problem in Eq. (1) has a unique solution and if so, under which conditions. However, this does not invalidate the proposed measures since only the reduction of the average travel time T is relevant.

A further improvement of the proposed method could be to use self-organized traffic light controls [Lam08]. This allows to restrict green times based on local measurements at downstream road segments. In the future, we aim to develop a new concept to model non-balanced traffic states. This includes a decision model for drivers facing long red times due to blockages on upstream road segments.

Acknowledgements

The authors thank the DFG (German Research Foundation) for partial financial support of this research and Gesche Roßfor the calibration and simulation of the Dresden network.

References

[Abu08] M. ABUELELA, S. OLARIU, and M. WEIGLE: "NOTICE: An Architecture for the Notification of Traffic Incidents". In: *Vehicular Technology Conference, 2008. VTC Spring 2008. IEEE*. 2008, pp. 3001–3005. DOI: 10.1109/VETECS.2008.320.

[Arn91] R. ARNOTT, A. de PALMA, and R. LINDSEY: "Does providing information to drivers reduce traffic congestion?" In: *Transportation Research Part A: General* 25.5 (1991), pp. 309–318. ISSN: 0191-2607.

[Boa10] T. R. BOARD: *Highway Capacity Manual*. Transportion Research Board. 2010.

[Dag07] C. F. DAGANZO: "Urban gridlock: Macroscopic modeling and mitigation approaches". In: *Transportation Research Part B: Methodological* 41.1 (2007), pp. 49–62.

[Dud75] C. L. DUDEK: "Better Management of Traffic Incidents: Scope of the Problem". In: *Transportation Research Board Special Report* 153 (1975), pp. 116–122.

[Gal89] A. I. GALL and F. L. HALL: "Distinguishing between incident congestion and recurrent congestion: a proposed logic". In: *Transportation Research Record* 1232 (1989), pp. 1–8.

[Ker05] B. S. KERNER, C. DEMIR, R. G. HERRTWICH, S. L. KLENOV, H. REHBORN, M. ALEKSIC, and A. HAUG: "Traffic state detection with floating car data in road networks". In: *Intelligent Transportation Systems, 2005. Proceedings. 2005 IEEE*. Sept. 2005, pp. 44–49.

[Lam08] S. LÄMMER and D. HELBING: "Self-Control of Traffic Lights and Vehicle Flows in Urban Road Networks". In: *Journal of Statistical Physics* (2008), P04019.

[Li08] M. LI: "Robustness Analysis for Road Networks". PhD thesis. Delft University of Technology, 2008.

[Rob69] D. I. ROBERTSON: "TRANSYT Method for Area Traffic Control". In: *Traffic Engineering & Control* 10 (1969), pp. 276–281.

[Sam00] A. SAMANT and H. ADELI: "Feature Extraction for Traffic Incident Detection Using Wavelet Transform and Linear Discriminant Analysis". In: *Computer-Aided Civil and Infrastructure Engineering* 15.4 (2000), pp. 241–250. ISSN: 1467-8667.

[Sch12] D. L. SCHRANK, B. EISELE, and T. J. LOMAX: *TTI's 2012 Urban Mobility Report*. Texas Transportation Institute, The Texas A&M University System, 2012.

[Sne10] M. SNELDER: "Designing Robust Road Networks". PhD thesis. Delft University of Technology, 2010.

[Tan05] S. TANG and H. GAO: "Traffic-incident detection-algorithm based on nonparametric regression". In: *Intelligent Transportation Systems, IEEE Transactions on* 6.1 (Mar. 2005), pp. 38–42. ISSN: 1524-9050.

[Tri00] M. TRIVEDI, I. MIKIC, and G. KOGUT: "Distributed video networks for incident detection and management". In: *Intelligent Transportation Systems, 2000. Proceedings. 2000 IEEE*. 2000, pp. 155–160.

[Uni00] UNITED STATES DEPARTMENT OF TRANSPORTATION: *Traffic Incident Management Handbook*. 2000.

[War52] J. G. WARDROP: "Some Theoretical Aspects of Road Traffic Research". In: *Proceedings of the Institute of Civil Engineers*. Vol. 2. 1. 1952, pp. 325–378.

[Wun98] K. WUNDERLICH: "A simulation-based assessment of route guidance benefits under variable network congestion conditions". In: *Mathematical and Computer Modelling* 27 (1998), pp. 87–101. ISSN: 0895-7177.

Corresponding author: Markus Rausch, Technische Universität Dresden, "Friedrich List" Faculty of Transportation and Traffic Sciences, 01062 Dresden, Germany, phone: +49 351 463 36878, e-mail: rausch@vwi.tu-dresden.de

Simulation Study of a Traffic Light Assistant based on Vehicle-Infrastructure Communication

Martin Treiber
Technische Universität Dresden

Abstract

Vehicle-infrastructure communication opens up new ways to improve traffic flow efficiency at signalized intersections. In this study, we assume that equipped vehicles can obtain information about switching times of relevant traffic lights in advance, and additionally counting data from upstream detectors. By means of simulation, we investigate, how equipped vehicles can make use of this information to improve traffic flow. Criteria include cycle-averaged capacity, driving comfort, fuel consumption, travel time, and the number of stops. Depending on the overall traffic demand and the penetration rate of equipped vehicles, we generally find several percent of improvement.

Keywords: Traffic light assistant, intelligent-driver model, adaptive cruise control, ACC, V2X, infrastructure-to-vehicle communication

1 Introduction

Individual vehicle-to-vehicle and vehicle-to-infrastructure communication, commonly referred to as V2X, are novel components of intelligent-traffic systems (ITS). Besides more traditional ITS applications such as variable speed limits on freeways or traffic-dependent signalization [Hun81; Low82], V2X promises new applications to make traffic flow more efficient or driving more comfortable and economic. While there are many investigations focussing on technical issues such as connectivity given a certain hop strategy, communication range, and percentage of equipped vehicles (penetration rate), e.g., [Thi08], few papers have investigated actual strategies to improve traffic flow characteristics. On freeways, a jam-warning system based on communications to and from road-side units (RSUs) has been proposed [Kra08]. Furthermore, a traffic-efficient adaptive-cruise control (ACC) has been proposed which relies on V2X communcation to determine the local traffic context influencing, in turn, the ACC parameterization [Kes10]. Regarding city traffic, early forms of V2X

have been investigated in the European projects Prometheous/Drive [Cat91]. However, these initiatives were more focussed on safety and routing information without explicitely treating any interactions with traffic lights. The investigation which is arguably most related to our work is the thesis [Ott11] on cooperative traffic control in cities discussing in depth the possibly destructive interplay between V2X (traffic-dependent signalization) and X2V (driver information relying on predetermined signalization).

In this contribution, we focus on city traffic at signalized intersections and investigate a set of strategies that is complementary to the self-controlled signal control strategy of Lämmer and Helbing [Lam08]: While, in the latter, the traffic lights "know" the future traffic, we assume that equipped vehicles know the future states of the next traffic light. In principle, the resulting traffic-light assistant (TLA) can operate in the information-based manual mode, or in the ACC-based automatic mode on which we will focus in this work.

In the next section, we lay out the methodology of this simulation-based study and define the objectives. Section 3 presents and analyzes the actual strategies "economic approach", "anticipative start", and "flying start". In the concluding Section 4, we discuss the results and point at conditions for implementing the strategies in an actual TLA.

2 Methodology

2.1 Car-Following Model

In order to get valid results, the underlying car-following model must be (i) sufficiently realistic to represent ACC driving in the automatic mode of the TLA, (ii) simple enough for calibration, and (iii) intuitive enough to readily implement the new strategies by re-parameterizing or augmenting the model. We apply the "Improved Intelligent-Driver Model" (IIDM) as described in Chapter 11 of the book [Tre13]. As the original Intelligent-Driver Model (IDM) [Tre00], it is a time-continuous car-following model with a smooth acceleration characteristics. Assuming speeds v not exceeding the desired speed v_0, its acceleration equation as a function of the (bumper-to-bumper) gap s, the own speed v and the speed v_l of the leader reads

$$\frac{\mathrm{d}v}{\mathrm{d}t} := a_{\mathrm{IIDM}}(s, v, v_l) = \begin{cases} (1 - z^2)\, a & z = \frac{s^*(v,v_l)}{s} \geq 1, \\ \left(1 - z^{\frac{2a}{a_{\mathrm{free}}}}\right) a_{\mathrm{free}} & \text{otherwise,} \end{cases} \tag{1}$$

where the expressions for the desired dynamic gap $s^*(v, v_l)$ and the free-flow acceleration $a_{\mathrm{free}}(v)$ are the same as that of the IDM,

$$s^*(v, v_l) = s_0 + \max\left[vT + \frac{v(v - v_l)}{2\sqrt{ab}},\, 0\right], \tag{2}$$

$$a_{\mathrm{free}}(v) = a\left[1 - \left(\frac{v}{v_0}\right)^{\delta}\right]. \tag{3}$$

The IIDM has the same parameter set as the IDM: desired speed v_0, desired time gap T, minimum space gap s_0, desired acceleration a, and desired deceleration b. However, it resolves two issues of the basic IDM when using it as an ACC acceleration controller: (i) the IIDM time-gap parameter T describes exactly the time gap in steady-state car-following situations while the actual IDM steady-state time gaps are somewhat larger [Tre00], (ii) a platoon of vehicle-drivers with same desired speed v_0 will not disperse over time as would be the case for the IDM. Notice that a slightly different formulation, the "IDM plus" with the acceleration function $a_{\mathrm{IDM+}}(s, v, v_l) = \min[a_{\mathrm{free}}, (1 - z^2)a]$, would serve this purpose as well.

By describing the vehicle motion with a time-continuous car-following model, we have neglected the in-vehicle control path since such models implicitely reflect an acceleration response time of zero. It might be necessary to explicitely model vehicle responses by a sub-microscopic model (e.g., PELOPS) when actually deploying such a system.

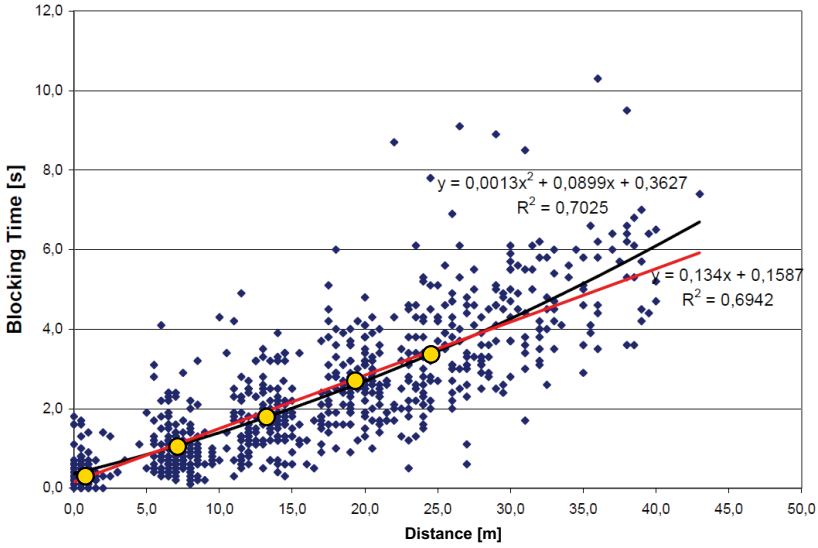


Figure 1: Calibration of the microscopic model with respect to the starting times and positions of a queue of waiting vehicles relative to the begin of the green phase (solid circles). Data are of the measurements in [Kuc08].

2.2 Calibration

Since we will investigate platoons travelling from traffic light to traffic light, the acceleration model parameters v_0, T, s_0, a, and b and the vehicle length l_{veh} (including their variances) should be calibrated to data of starting and stopping situations.

For calibrating a, T, and the combination $l_{\mathrm{eff}} = l_{\mathrm{veh}} + s_0$ (effective vehicle length), we use the empirical results of Kücking [Kuc08] taken at three intersections in the city of Hannover, Germany. There, the "blocking time" of the n^{th} vehicle of a waiting queue (the time interval this vehicle remains stopped after the light has turned green) has been measured vs. the distance of this vehicle to the stopping line of the traffic light. Figure 1 reproduces these data together with the simulation results (orange bullets) for the calibrated parameters $l_{\mathrm{eff}} = 6.5\,\mathrm{m}$,

$a = 1.5\,\mathrm{m/s^2}$, and $T = 1.2\,\mathrm{s}$ assuming identical driver-vehicles. Further simulations with heterogeneous drivers and vehicles reveal that independently and uniformly distributed values for l_{eff}, T and a with standard deviations of the order of 30 % of the respective expectation value can reproduce the observed data scatter and its increase with the vehicle position (for positions $n = 5$ and higher, the scattering does no longer allow to identify n). Moreover, since trucks are excluded from the measurements, it is reasonable to assume that the observed cars have an average length of 4.5 m resulting in an expectation value $s_0 = 2\,\mathrm{m}$ for s_0.

For estimating the comfortable deceleration, the approach to a red traffic light is relevant. Investigations on the Lankershim data set of the NGSIM data [NGS12] including such situations show that a typical deceleration is $b = 2\,\mathrm{m/s^2}$ [Vit10]. Finally, for the desired speed, we assumed a fixed value of $v_0 = 50\,\mathrm{km/h}$ representing the usual inner-city speed limit in Germany.

2.3 Simulation

While the parameters clearly are distributed due to inter-vehicle and inter-driver variations, it is nevertheless necessary to use the same vehicle population for all the following simulation experiments. Specifically, we use following sequence of four vehicle-driver combinations: 1. average driver (expectation values for the parameters), 2. agile driver (a increased to $2\,\mathrm{m/s^2}$, $T = 1.8\,\mathrm{s}$), 3. less agile but anticipative driver (a and b decreased to $1.2\,\mathrm{m/s^2}$ and $1\,\mathrm{m/s^2}$, respectively), and 4. a truck ($l_{\mathrm{veh}} = 12\,\mathrm{m}$, $T = 1.7\,\mathrm{s}$, and $a = b = 1\,\mathrm{m/s^2}$). If necessary, this sequence is repeated. Figure 2 shows the simulation result for the start-and-stop reference scenario against which the strategies of the traffic light assistent will be tested in the next section.

2.4 Traffic Flow Metrics

In the ideal case, the TLA reduces the travel time of the equipped and the other vehicles, increases driving comfort and traffic flow efficiency, and reduces fuel consumption. To assess travel time, we use the average speed of a vehicle, or average over all vehicles during the complete simulation run. As proxy for the driving comfort, we take the number of stops during one simulation, or, equivalently, the fraction of stopped vehicles. Traffic flow efficiency is equivalent to the cycle-averaged dynamic capacity, i.e., the average number of vehicles passing a traffic light per cycle in congested congestions in the absence of gridlocks. Finally, we determine the fuel consumption by a physics-based modal consumption model as described in Chapter 20.4 of the book [Tre13]). Such models take the simulated trajectories and some vehicle attributes as input and return the instantaneous consumption rate and the total consumption of a given vehicle. To be specific, we assume a mid-size car with following attributes: Characteristic map of a 118 kW gasoline engine as in Fig. 20.4 of [Tre13], idling power $P_0 = 3\,\mathrm{kW}$, total mass $m = 1\,500\,\mathrm{kg}$, friction coefficient $\mu = 0.015$, air-drag coefficient $c_d = 0.32$, frontal cross-section $A = 2\,\mathrm{m^2}$, a dynamic tyre radius $r_{\mathrm{dyn}} = 0.286\,\mathrm{m}$. Furthermore, we assume a five-gear transmission with transmission ratios of 13.90, 7.80, 5.25, 3.79, and

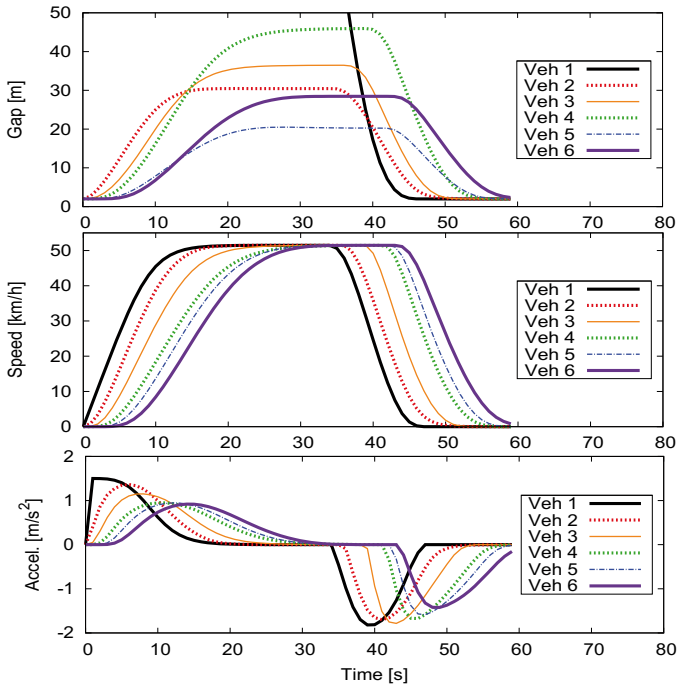


Figure 2: Start and stop of the simulated platoon of heterogeneous vehicle-drivers in the reference case.

3.09, respectively, and chose the most economic gear for a given driving mode characterized by v and $\frac{\mathrm{d}v}{\mathrm{d}t}$. The engine power management includes overrun-fuel cutoff, idling when the vehicle is stopped, and no energy recuperation during braking.

3 Strategies of the Traffic Light Assistant and their Simulation

The appropriate TLA strategy depends essentially on the arrival time at the next traffic light relative to its phases. We distinguish following approaching situations (cf. Fig. 3):

- A stop is unavoidable (the first seven of the red trajectories of Fig. 3),
- anticipative start compensating for the reaction time of the first vehicle (last red trajectory),
- flying start realized by anticipative braking (blue trajectories),
- free passage (green trajectories),
- temporary “boost” to catch the last part of the green phase (violet trajectories).

Nothing can be done in the situation of a free passage while the “boost” strategy implies temporarily exceeding the speed limit. Therefore, we will only develop and simulate strategies for the first three situations. Generalizing the above sketch, we will also investigate how other (equipped or non-equipped) vehicles will affect the strategies. Furthermore, by a complex simulation over several cycles, we investigate any (positive or negative) interactions between the strategies and between equipped and non-equipped vehicles.

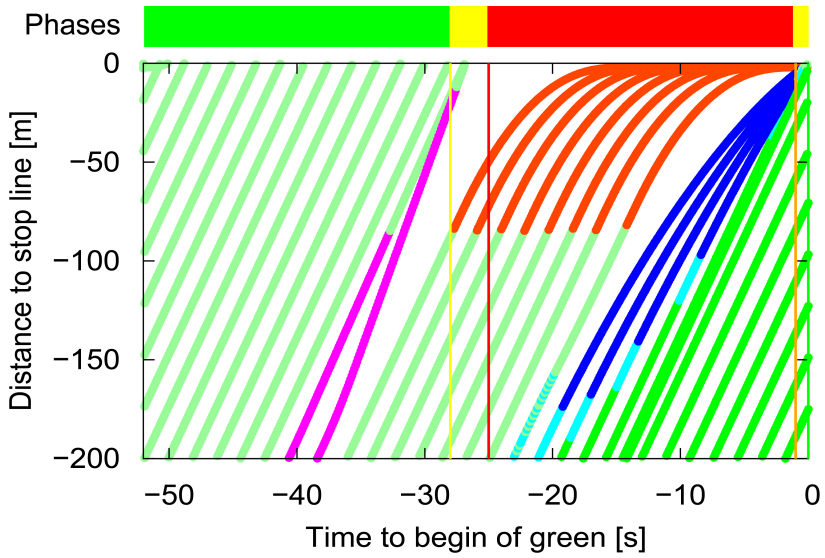


Figure 3: Approach situations relative to the phases of the traffic light. Each trajectory corresponds to an individual simulation of the considered vehicle with no interactions to other vehicles. For the color coding, see the main text.

3.1 Approach to a Stop

In certain situations, a stop behind a red light or a waiting queue is unavoidable. This situation is true if (i) extrapolated constand-speed arrival occurs during a red phase, and (ii) the "flying-start" strategy of Sect. 3.3 would produce minimum speeds below a certain threshold which we assumed to be $v_{\min}^{\text{flying}} = 10\,\mathrm{km/h}$. Notice that this szenario may also apply for approaching green traffic lights if the car cannot make it to the traffic light before switching time: In such a situation, drivers of non-equipped cars would just go ahead braking later and necessarily harder. While this situation is not relevant for improving flow efficiency, it is nevertheless possible to reduce fuel consumption by early use of the engine brake, i.e., early activation of the overrun cut-off.

In the car-following model, we implement this strategy by reducing the comfortable deceleration from $b = 2\,\mathrm{m/s^2}$ to $1\,\mathrm{m/s^2}$ (homogeneous driver-vehicle population), or by 50 % for each vehicle (heterogeneous population). Reducing the desired deceleration means earlier braking, in line with this strategy.

Figure 4 shows speed and consumption profiles for an equipped vehicle (solid lines) vs. the reference (dotted). The equipped vehicle itself saves about 3.5 ml of fuel (6 % for the complete start-stop cycle). The two next (non-equipped) followers save about 3 % and 1 %, respectively.

3.2 Anticipative Start

The rationale of the strategy of the anticipative start is to compensate for the reaction time delay τ. Since the reaction time is only relevant for the driver of the first vehicle in a queue,

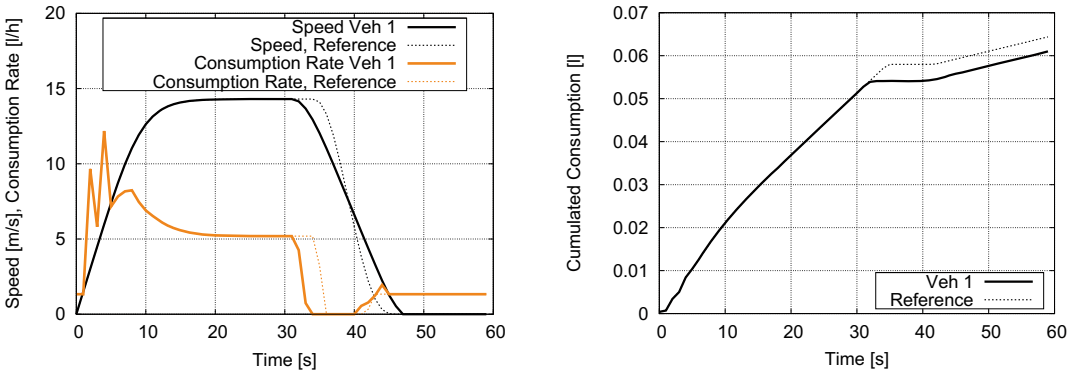


Figure 4: Fuel-saving approach to a waiting queue. Left: speed profile and instantaneous consumption rate; right: cumulative consumption during the complete start-stop cycle.

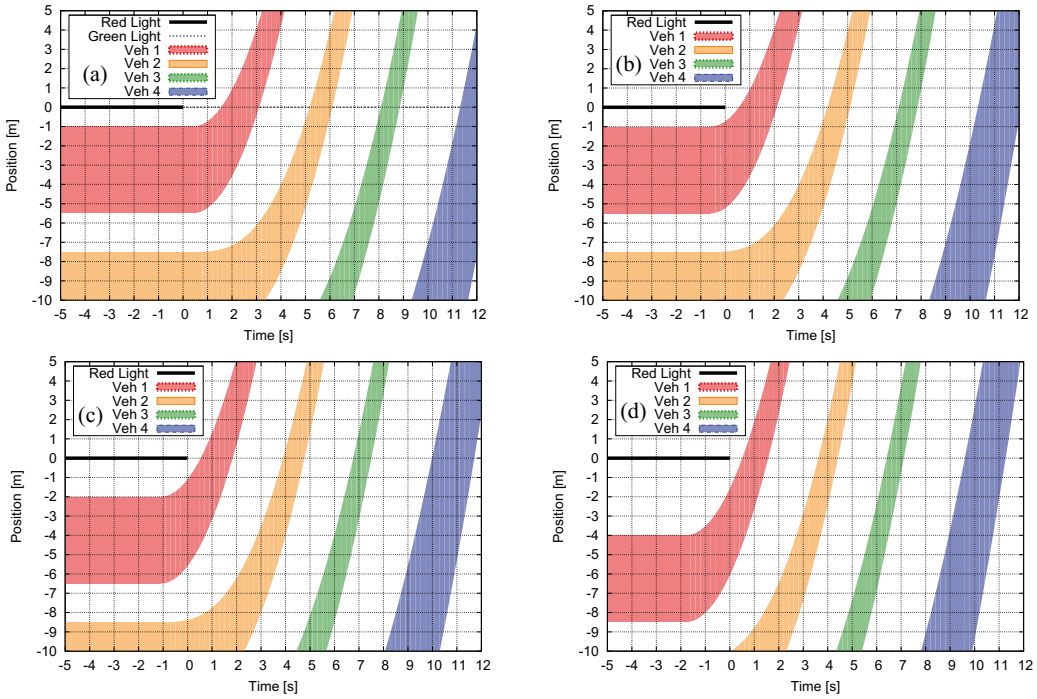


Figure 5: Start at green from the first position of a queue of waiting vehicles. (a) reference; (b) anticipative start; (c) anticipative start plus 1 m additional gap; (d) anticipative start plus 3 m additional gap.

the anticipative-start strategy is restricted to this vehicle. In the reference case corresponding to the calibrated parameters (Fig. 5 (a)), the front of the first vehicle crosses the stopping line about 1.5 s after the change to green corresponding to $\tau \approx 0.7\,\mathrm{s}$ (the rest of the time is needed to move the first meter to the stopping line). If this vehicle started one second earlier, i.e., before the switching to green (Fig. 5 (b)), the situation is yet save but an average of 0.5 additional vehicles can pass during one green phase assuming an outflow of 1 800 veh/h after some vehicles. Considering the 12 vehicles that would pass in the reference scenario

during the 30 s long green phase of the 60 s cycle, this amounts, on average, to an increase by 4 %. An additional second can be saved, allowing 13 instead of 12 vehicles per green phase, if the first vehicle stops 4 m upstream of the stopping line (instead of 1 m) allowing an even earlier start without compromising the safety (Fig. 5(d)). However, there are limits in terms of acceptance and available space, so stopping 2 m before the stopping line (Fig. 5(c)) is more realistic. In effect, the latter strategy variants transform the anticipative start in a "flying start" which we will discuss now.

3.3 Flying Start

If, relative to the phases, a vehicle arrives later than in the previous two situations but too early to have a free passage, preemptive braking may avoid a stop or, at least, increase the minimum speed during the approaching phase. As depicted in Fig. 6, the strategy consists in controlling the vehicles's ACC such that a certain spatiotemporal *target point* $(\Delta x, \Delta t)$ relative to the stopping line and the switching time to green is reached. This point is determined such that a minimum of speed reduction is realized without impairing traffic efficiency by detaching this vehicle from the platoon of leaders.

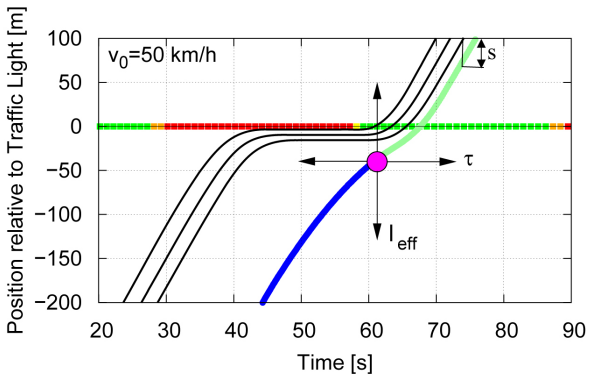


Figure 6: Preemptive braking to avoid a stop: Spatiotemporal target for the 4^{th} vehicle (pink circle). The arrows indicate how the target changes when varying the reaction time τ or the effective length l_{eff}.

From basic kinematic theory [Lig55] and the properties of the IDM it follows that the propagation velocity c of the *positions* of the vehicles at the respective starting times is constant and given by $c \approx -l_{\text{eff}}/\tilde{T}$ where $\tilde{T}$ is of the order of the IDM parameter T. Assuming a gap s_0^* of the first waiting vehicle to the stopping line and a reaction delay τ of its driver, the estimated spatiotemporal *starting point* of the n^{th} vehicle reads

$$(\Delta x, \Delta t) = (s_0^* + (n-1)\, l_{\text{eff}}, \tau + (n-1)\, \tilde{T}). \tag{4}$$

The points lie on a straight line which is consistent with observations (filled circles in Fig. 1). While we assume that, by additional V2X communication from a stationary detector to the

vehicle, the equipped vehicle knows its order number n, there are uncertainties in τ, l_{eff}, and T which depend on unknown properties of the vehicles and drivers ahead. Furthermore, since the strategy tries to avoid a stop, the *target* point lies several meters upstream of and/or a few seconds after the anticipated starting point.

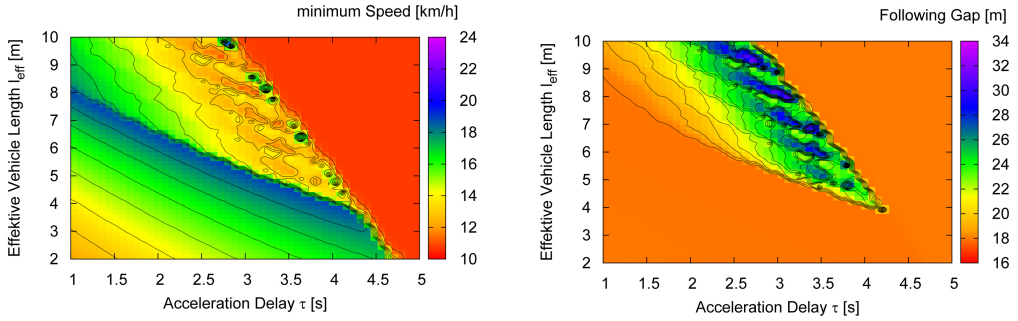


Figure 7: Robustness of the premptive braking strategy. Shown is its efficiency for the 3^{rd} vehicle in terms of the minimum speed during the approach (left) and the gap once this vehicle is 50 m downstream of the traffic light (right).

Is this strategy nevertheless robust? In order to assess this, we treat τ and l_{eff} as free parameters of (4) to be estimated and plot the performance metrics minimum speed $v_{\min}$ characterizing driving comfort and spatial gap s to the platoon (cf. Fig. 6) characterizing the dynamic capacity as a function of τ and l_{eff}.

Figure 7 shows these metrics for the $n = 3^{\text{rd}}$ vehicle arriving at a timing such that the minimum speed would be $v_{\min} = 10\,\text{km/h}$ if this vehicle were not equipped. For the best estimates (e.g., $l_{\text{eff}} = 6.5\,\text{m}$ and $\tau = 2\,\text{s}$), this minimum speed is nearly doubled without compromising the capacity which would be indicated by an increased following gap s. The simulations also show that estimation errors have one of three consequences: (i) if the queue length and dissolution time are estimated too optimistically (l_{eff} and τ too small), there is still a positive effect since the minimum speed is increased without jeopardizing the efficiency; (ii) if the queue is massively overestimated (l_{eff} and τ significantly too large), the whole strategy is deemed unfeasible and the approach reverts to that of non-equipped vehicles; (iii) if, however, the queue is only slightly overestimated, the strategy kicks in ($v_{\min}$ increases) but the capacity is reduced since s increases as well: the car does no longer catch the platoon. A look at the parameter ranges (the plots range over factors of five in both τ and l_{eff}) indicates that this strategy is robust when erring on the optimistic side, if there is any doubt.

Finally, we mention that counting errors (e.g. due to a vehicle changing lanes when approaching a red traffic light meaning that this vehicle has not passed the correct stationary detector) will lead to similar errors for the estimated target point as above. Consequently, this strategy should be robust with respect to counting errors as well.

3.4 Complex Simulation

In the previous sctions, we have investigated the different strategies of the TLA in isolation. However, there are interactions. For example, the optimal target point of the flying-start strategy is shifted backwards in time when equipped leading vehicles apply the anticipative-start strategy. Furthermore, the question remains if the TLA remains effective if there is significant surrounding traffic (up to the level of saturation) and whether the results are sensitive to the order in which slow and fast, equipped and non-equipped vehicles arrive.

We investigate this by complex simulations of all strategies over several cycles where we vary, in each simulation, the overall traffic demand (inflow) Q_{in}, and the penetration rate p of equipped vehicles. Unlike the simulations of single strategies, we allow for full stochasticity in the vehicle composition. At inflow, we draw, for each new vehicle, the model parameters from the independent uniform distributions specified in Sect. 2.3 and assign, with a probability p, the property “is equipped”.

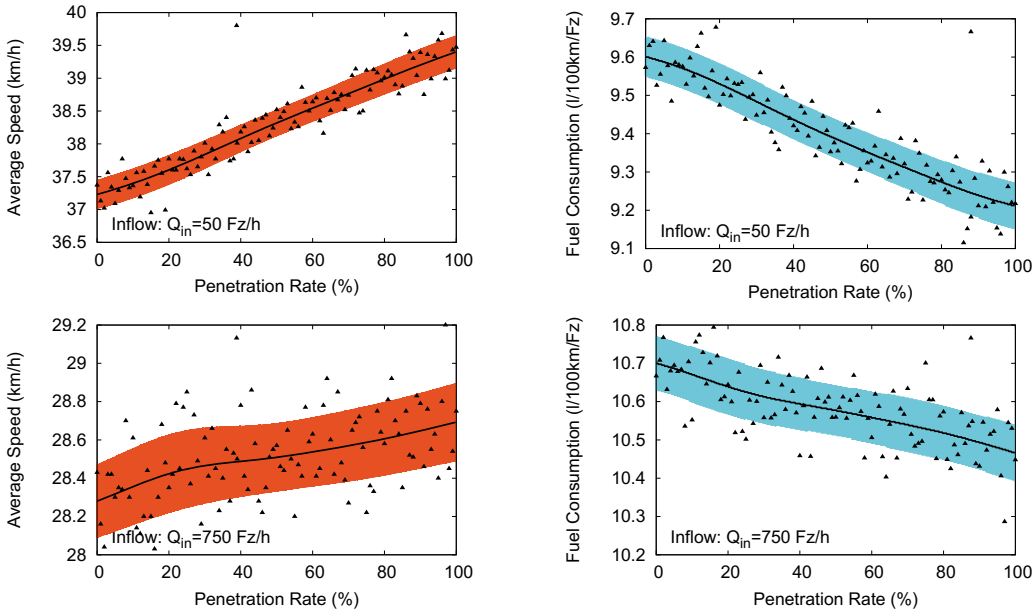


Figure 8: Complex simulation of the overall effectiveness for all vehicles over several cycles (see the main text for details).

Figure 8 shows the results for the performance metrics “average speed” (which is related to the average travel time), and “average consumption” as a function of the penetration rate for a small traffic demand (top row), and a demand near saturation (bottom). Each symbol corresponds to a simulation for given values of Q_{in} and p. Due to the many stochastic factors and interactions, we observe a wide scattering. Determining the local average (solid lines) and $\pm 1\sigma$ bands (colored areas) by kernel-based regression (kernel width 15 %), we nevertheless detect significant systematic effects. For low traffic demand, we observe that both travel times and fuel consumption are reduced by about 4 % when going from the reference to $p = 100\,\%$ penetration. Furthermore, the effects essentially increase linearly

with p, so the *relative performance indexes* I_{T} and I_{C} with respect to travel time T_t and fuel consumption C,

$$I_{\mathrm{T_t}} = -\frac{1}{T_t}\frac{\partial T_t}{\partial p}, \quad I_{\mathrm{C}} = -\frac{1}{C}\frac{\partial C}{\partial p} \tag{5}$$

are both constant and of the order of 4 %. Similar performance indexes are obtained for the performance metrics "number of stops". For higher traffic demand (lower row of Fig. 8), the relative performance of the TLA decreases except for the metrics "dynamic capacity".

4 Discussion

We have investigated, by means of simulation, a concept of a traffic-light assistant (TLA) containing three strategies to optimize the approach to and starting from traffic lights: "economic approach", "anticipative start", and "flying start". The strategies are based on V2X communication: In order to implement the TLA, equipped vehicles must obtain switching information of the relevant traffic lights and – as in the self-controlled signal strategy [Lam08] – counting data from a detector at least 100 m upstream of the traffic light. Complex simulations including all interactions show that, for comparatively low traffic demand, the TLA is effective. To quantify this, we introduced relative performance indexes which we consider to be the most universal approach to assess penetration effects of individual-vehicle based ITS. For our specific setting (maximum speed 50 km/h, cycle time 60 s, green time 30 s), we obtained performance indexes of about 4 % for most metrics if traffic demand is low. We obtain higher values for higher maximum speeds and lower cycle times, and lower values for a higher demand. While the relative performance is generally lower than that of the traffic-adaptive ACC on freeways (about 25 %) [Kes10], the *individual* advantage kicks in with the first equipped vehicle, in contrast to traffic-adaptive ACC.

Acknowledgements

We would like to thank the Volkswagen AG who has sponsored part of this work in a project.

References

[Cat91] I. CATLING and B. MCQUEEN: "Road transport informatics in Europe-major programs and demonstrations". In: *Vehicular Technology, IEEE Transactions on* 40.1 (1991), pp. 132–140.

[Hun81] P. HUNT, D. ROBERTSON, R. BRETHERTON, and R. WINTON: *SCOOT-a traffic responsive method of coordinating signals*. Tech. rep. 1981.

[Kes10] A. KESTING, M. TREIBER, and D. HELBING: "Enhanced Intelligent Driver Model to access the impact of driving strategies on traffic capacity". In: *Philosophical Transactions of the Royal Society A* 368 (2010), pp. 4585–4605.

[Kra08] F. KRANKE and H. POPPE: "Traffic Guard - Merging sensor data and C2I/C2C information for proactive congestion avoiding driver assistance systems". In: *FISITA World Automotive Congress*. 2008.

[Kuc08] KÜCKING: *Analyse des Verkehrsablaufs an signalisierten Kreuzungen - wie schnell lösen sich Rückstaus auf?* Volkswagen AG, unpublished. 2008.

[Lam08] S. LÄMMER and D. HELBING: "Self-control of traffic lights and vehicle flows in urban road networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.04 (2008), P04019.

[Lig55] M. J. LIGHTHILL and G. B. WHITHAM: "On Kinematic Waves. II. A Theory of Traffic Flow on Long Crowded Roads". In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 229.1178 (1955), pp. 317–345. DOI: 10.1098/rspa.1955.0089. URL: http://rspa.royalsocietypublishing.org/content/229/1178/317.full.pdf.

[Low82] P. LOWRIE: "The Sydney coordinated adaptive traffic system-principles, methodology, algorithms". In: *International Conference on Road Traffic Signalling, 1982, London, United Kingdom*. 207. 1982.

[NGS12] *NGSIM - Next Generation Simulation*. (Last Access 14 August 2012). URL: http://ngsim-community.org/.

[Ott11] T. OTTO: *Kooperative Verkehrsbeeinflussung und Verkehrssteuerung an signalisierten Knotenpunkten*. Vol. 21. kassel university press GmbH, 2011.

[Thi08] C. THIEMANN, M. TREIBER, and A. KESTING: "Longitudinal hopping in inter-vehicle communication: Theory and simulations on modeled and empirical trajectory data". In: *Phyical Review E* 78 (2008), p. 036102.

[Tre00] M. TREIBER, A. HENNECKE, and D. HELBING: "Congested traffic states in empirical observations and microscopic simulations". In: *Physical Review E* 62 (2000), pp. 1805–1824.

[Tre13] M. TREIBER and A. KESTING: *Traffic Flow Dynamics: Data, Models and Simulation*. Berlin: Springer, 2013.

[Vit10] F. VITI, S. P. HOOGENDOORN, H. J. van ZUYLEN, I. R. WILMINK, and B. VAN AREM: "Microscopic data for analyzing driving behavior at traffic signals". In: *Traffic Data Collection and its Standardization*. Springer, 2010, pp. 171–191.

Corresponding author: Martin Treiber, Technische Universität Dresden, "Friedrich List" Faculty of Transportation and Traffic Sciences, Institute of Traffic Economics, phone: +49 351 463 36794, e-mail: treiber@vwi.tu-dresden.de

Vehicular Traffic Monitoring through VANETs: Simulation and Analysis in a Real Case Study

Andrea Baiocchi[1], Chiara Colombaroni[2], Francesca Cuomo[1], Mario De Felice[1], Gaetano Fusco[1]
[1]University of Roma "La Sapienza"
[2]University of Roma "Niccolò Cusano"

Abstract

The paper deals with application of VANETs technology to road traffic monitoring and presents a distributed communication protocol, aiming at vehicular traffic monitoring. A realistic simulation of a main expressway in Roma, Italy, calibrated by exploiting information provided by a significant sample of floating car data, allows to individuate the operational limits of the communication protocols and to evaluate the accuracy of the estimates of traffic state.

Keywords: Traffic measurement, VANET, dissemination protocol, trace driven traffic simulation.

1 Introduction

Vehicular ad hoc networks (VANETs) are the technology for building wireless networks among mobile vehicles equipped with On Board Units (OBU), and including also fixed Road Side Units (RSU). Thanks to the Dedicated Short Range Communications (DSRC) standards like Wireless Access in Vehicular Environments (WAVE) [Mor10] vehicles are able to exchange data messages for accessing both safety and infotainment applications.

VANETs constitute a diffuse, autonomous, scalable network that can perform distributed traffic monitoring and traveler information together. VANETs can support Intelligent Transportation Systems (ITS) with both Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) communications. Not only traffic monitoring and communication tasks are transferred to vehicle communication devices. Also a significant part of data processing can be distributed and conveyed to the vehicle communication network, since on-board devices can do in-network processing with data coming from other vehicles, without requiring that each single vehicle communicate with a traffic control centre.

This work defines a VANET based road traffic monitoring system, based on a distributed communication protocol. To gauge the feasibility of the approach, we have set up a realistic simulation of a main expressway in Roma, Italy, calibrated by exploiting information provided by an extensive sample of floating car data. Joint vehicular mobility and communication network simulations have been carried out, pointing out that the designed protocol can collect quite accurate average speed estimates for different section of the road with a very limited effort. No infrastructure is needed, except a single RSU for a road span of almost 70 km, collection times are within few seconds and relative errors range between 4% and 5.6% on the average.

The rest of the paper gives a short survey of related works (Section 2), then the VANET communication and monitoring protocol is defined (Section 3). The considered test case and the performance evaluation are presented in Section 4 and 5. respectively. Final remarks are given in Section 6.

2 Related Work

VANETs provide a distributed in-vehicle technology for localization-based systems [Ban01] and many authors studied their possible applications to different kinds of ITS. Applications on active safety ranged from danger warning system [Mit10], Cooperative Adaptive Cruise Control [Kia12], and cooperative driving [Sep13]. Applications of cooperative collision warning at intersection approaching allow integrating Advanced Driver Assistance (ADAS) and Advanced Traffic Management Systems (ATMS) and implementing a sort of virtual signal, as firstly proposed by [Hua04]. Further enhancements were later introduced, among others, by [Naf12][Far12]. [Nas12] studied an intersection traffic signalling system and evaluated its stability properties by applying the microsimulation model SUMO [Beh11] to an isolated road intersection. [Far12] presented one of first on the field experiment of an infrastructure-to-vehicle system aimed at assessing the impacts of cooperative driving on users' behaviour under real traffic conditions on a 17-km long motorway segment in Austria.

[Ger12] highlighted the need for careful evaluation of VANET by simulation models before large scale deployments. Several software packages have been developed to evaluate VANET performances in simulation environment. A review is provided by [Zea12]. Several authors studied the effectiveness of VANET inter-vehicle communications to recognize traffic conditions by using probe vehicle information and evaluated the performances of the communication system by simulation on simple idealistic test cases, with the possible exception of [Leo11], who carried out a simulation of a VANET-based ITS systems in the downtown area of Portland, Oregon (USA). [Che12] tested connectivity among vehicles by analyzing different inter-vehicle spacing probability distributions. [Qiu12] remarked that most of the previous network performance models paid little attention to vehicle distribution, or simply assumed homogeneous car distribution. Based on the knowledge of car density at each location from the traffic model, they developed an IEEE 802.11p broadcasting model and introduced a new metric to better characterize the broadcasting performance and packet collision probability

in VANETs. [Yin13] proposed analytical models for the vehicle connectivity on two parallel roadways, by assuming general distributions for vehicle headways.

[Som11] argued that the mobility models used in many network simulation tools do not take into account driver behaviour or specific characteristics of the urban environment and highlighted the need for bidirectional coupling of realistic mobility models with network simulation tools in evaluations of VANET protocols. They applied a hybrid simulation framework composed of the network simulator OMNeT++ and the road traffic simulator SUMO and tested several incident scenarios in a realistic urban road network. Unlike the road network, traffic scenario was purely hypothetical and consisted in a simulation of 200 cars leaving a parking lot, on average one every 6 s. We agree with [Qiu12] and [Som11] arguments. So, we tackled the simulation of VANET monitoring system by building a realistic test case, on a 68 km long ring road expressway in Roma, Italy, which was calibrated through about 50,000 traces of GPS equipped vehicles, which provided a detailed picture of the actual traffic conditions.

3 The VANET Monitoring System

Several solutions have been proposed for message dissemination in a VANET [Pan12], with the goal of extending the coverage area reached by message flows originating from a given node, thanks to vehicle-to-vehicle multi-hop communications. A key issue is to avoid the broadcast storm problem [Ni99]. An effective distributed approach has been introduced in [Sun00], the so called Distance Defer Transfer (DDT). The basic principle is that the forwarding vehicle is picked as the one farthest away from the sender among all those vehicles that have received the message to be forwarded. To do that, each vehicle that receives a new message waits for a defer timer that decreases with the sender-receiver distance before retransmitting the message (*forwarding rule*, $\mathcal{FR}$). Moreover, a vehicle receiving multiple copies of a message is inhibited from forwarding it (*inhibition rule*, $\mathcal{IR}$). To this end, GPS positions are assumed to be available to nodes, which is plausible in the VANET case.

The dissemination protocol can be implemented as a separate layer from MAC, sitting on top of it, or it can be merged with the MAC protocol itself. The latter approach implies modifying the MAC logic and designing new firmware, so inter-operability with legacy MAC versions should be carefully tackled as well. Defining an independent upper layer, called *Forwarding Layer* (FL) yields flexibility and decouples the design of MAC from dissemination logic. This is useful since the MAC protocol can also serve different traffic flows, not necessarily to be broadcast for dissemination.

Given that the dissemination logic is implemented in the FL on top of MAC, we must account for the effect of a non ideal MAC protocol. With IEEE 802.11p CSMA/CA MAC protocol, if messages are issued well separated in time, so that each message propagates through the VANET without interacting with previous and subsequent messages, then the MAC service time is upper bounded by a constant value $\theta = DIFS + \sigma W_0 + T_{MAC}$, where $DIFS$ is the DCF Inter-Frame Spacing of IEEE 802.11p, σ is the back off slot, W_0 is the basic

contention window size and T_{MAC} is the time required to send out a MAC frame (it depends on the frame payload length and on the air bit rate). As a matter of example, with payload length of $1000\ bytes$ and air bit rate of $6\ Mbps$, it is $\theta \approx 1.7\ ms$.

Then, once the FL timer of a vehicle A expires, say at time t_0, the message moves down to the MAC layer entity (the network interface card) to be sent out on the radio channel. Vehicles within range of A cannot receive the message from A until time $t_0 + \theta$, due to the processing of the MAC protocol and of the physical layer. If the timer of another nearby vehicle B expires after t_0 and *before* time $t_0 + \theta$, the FL entity of B will not be inhibited, since B has not received the message from A yet. As a consequence, B as well will commit its MAC entity to sending a copy of the message. This additional, undesired forwarding actions are referred to as *spurious forwarding* and have first been noticed in [Sal13].

Spurious forwarding stops the packet dissemination. This catastrophic effect is triggered since vehicles that receive both the first forwarded message and the subsequent spuriously forwarded copies of the same message will apply the $\mathcal{IR}$. This harmful effect can be avoided by means of a *hop count* field in the message header. Let us denote as h this hop count number. A vehicle receiving a new message with sequence number k and hop count h, will schedule the forwarding of that message, by changing the hop count to $h + 1$. While this message is pending, waiting for the FL defer timer to expire, if A receives a message with the same sequence number k, the newly received message is considered a duplicate only if it has also hop count equal to $h + 1$. In that case, the node applies the $\mathcal{IR}$ and drops its copy of the message. If instead, the received message has sequence number k, but hop count h, that one is not considered as a duplicate and it does not trigger the $\mathcal{IR}$, since it is from a spurious forwarder of the previous hop zone.

Leveraging on the DDT idea, but accounting for the spurious forwarding effect, we define a so called Distance Based Forwarding (DBF) dissemination protocol. DBF operations is based on the Forwarding and Inhibition Rules. Let A be a vehicle located at a point of coordinates P_A, forwarding a message with sequence number k and hop count h, at time t. Any other vehicle V, at position P_V within range of A, receives the message sent by A. Let k_V the biggest message sequence number already seen and completely dealt with by V.

- *Forwarding Rule*. By checking that $k \leq k_V$, V can discard old or duplicated messages. If the message, is new, V schedules its forwarding ($\mathcal{FR}$), by setting a timer $T_{V,k} = T_{max}(1 - \overline{P_V P_A}/R_{max})$, where T_{max} is the maximum forwarding delay, R_{max} is an upper bound of the coverage radius of OBU transceivers, and $\overline{P_V P_A}$ is the distance between V and A. Hence, V schedules the forwarding of the message at time $t + T_{V,k}$ with sequence number k and hop count $h + 1$.

- *Inhibition Rule*. If during the time interval $(t, t + T_{V,k}]$, V receives another message with sequence number k and hop count h', V checks that $h' = h + 1$. In that case, V drops the scheduled message and will not forward it. Otherwise, no inhibition takes place.

Figure 1 shows an example where A sends a message with $h = 1$ and both B and C receive it. Since C is farther from A, its FL timer will expire first, then C will end up sending

the message first, thus inhibiting B. Hence, D and E receive the message forwarded by C, with $h = 2$. They both set their timers. Since the difference $\overline{P_D P_C} - \overline{P_E P_C}$ is too small, the difference between the two timer values T_D and T_E falls below θ, and both D and E will commit their radio interface equipment to sending a frame containing an instance of the same message, both of them with $h = 3$. Node F receives those two duplicated messages one after the other. The first received one is scheduled for forwarding with $h = 4$, while the second received one is discarded with no harm, thanks to its hop count value ($h = 3$) being smaller than the one scheduled by F ($h = 4$).

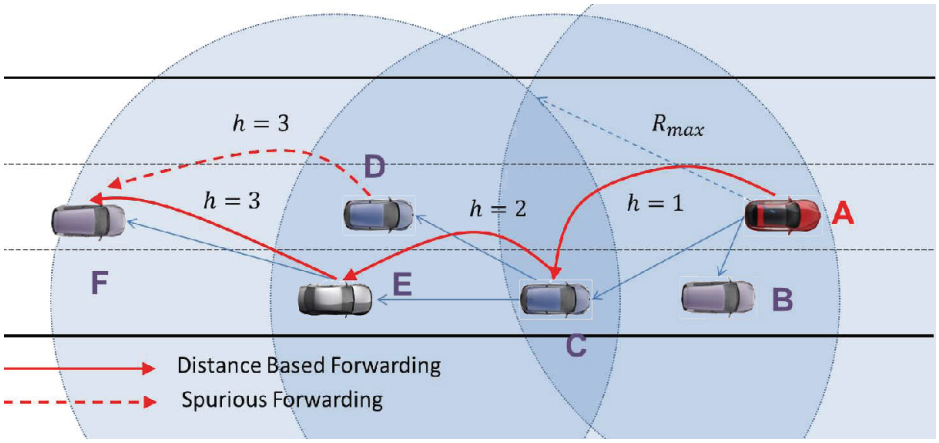


Figure 1: An example of distance based forwarding: the sequence of forwarding vehicles A-C-E-F is the ideal one. A spurious forwarding occur with node D.

We consider two Road Side Units (RSUs) located along a road span, RSU$_a$ and RSU$_b$. RSU$_a$ originates a stream of messages, issuing one *call for measurement collection* (cmc) message every time interval T_m. The cmc message is passed over from vehicle to vehicle by using a DBF logic, until it reaches RSU$_b$, that is the final sink of the collected measurements.

The logic defined to monitor the traffic on the road with the VANET is implemented by having a distributed collection of a vector $\mathbf{m} = (m_1, \ldots, m_n)$ where the ith element is a 3-tuple $m_i = \langle v_i, P_i, t_i \rangle$. Here v_i is the current speed of i-th sampled vehicle, P_i its geographical coordinates and t_i a timestamp. The size of $\mathbf{m}$ depends on the number of vehicles that collect this information, that is the number of vehicles forwarding the message (or the number of hops) from the source RSU up to the sink RSU. The message initially issued by RSU$_a$ has an empty payload. Each sampled vehicle, namely a vehicle acting as a forwarding node according to DBF, appends its 3-tuple to the current payload of the message. When the message reaches RSU$_b$, it carries all n collected measurements. n is related to the average hop length and to the length of the monitored road span. A fast polling is possible, since each forwarding hop takes a time in the order of ms and typical hop lengths can be several hundred of meters. The traveling speed of the monitoring messages can be thus in the order of 50 km/s, three orders of magnitude more than vehicle speed.

4 Description Of A Real-Size Test Case

To set up a realistic simulation, we analyzed a large database of about 104 millions of GPS traces collected by about 80,000 equipped vehicles that made about 9 millions trips during the month of May 2010 in the metropolitan area of Roma (Italy). We have selected a subset of 50,220 vehicle traces on the Ring Road (GRA), the 68 km long ring-shaped expressway surrounding the city of Roma, which collects and distributes long-haul traffic entering and exiting from the city. Each vehicle sends its GPS trace record every 30 seconds. A variety of information was captured in each record, including the record ID, vehicle ID, geographical coordinates, speed, quality of GPS signal.

The first step of our analysis was based on the segmentation of the GRA in 29 different segments of length L_j, $j = 1, \ldots, 29$, where the main exits from the GRA highway are the starting and ending points of each segment. Then vehicles were divided in two sets according to their traffic direction: clockwise and counterclockwise. Four different time periods of four hours each have been considered, starting from 7 am until 11 pm. Inter-vehicle distance distribution and speed distribution were obtained for each of the four time periods. This analysis showed that the highest density of vehicles is in the time period between 3 pm and 7 pm, which has the largest number of detected vehicles (9732 vehicles).

To set up a realistic mobility simulation of the urban traffic in the GRA the available sample of data have been inferred to the universe of vehicles by assuming a random uniform sampling. Let Δt be the sampling interval (30 s in our study), v_i the average speed of vehicle i, n_j the estimated number of vehicles traveling in the j-th segment, $g_j(t_1, t_2)$ the number of detected GPS signals in the j-th segment during the observation time interval $[t_1, t_2]$, L_j the length of the j-th segment, a the probe vehicles penetration rate ($a \approx 2.3\%$ in our study) and q_j the estimated flow on the j-th segment. Then $n_j = \sum_{i=1}^{g_j(t_1,t_2)} \frac{v_i \Delta t}{L_j}$ and the resulting flow is $q_j = \frac{n_j}{a(t_2 - t_1)}$. The above inference relations have been applied to the traffic flows in the peak period and have been used to enter into each segment j a flow of vehicles with intensity q_j. The flows q_j have been used to estimate the OD matrix, where Origins and Destinations correspond to the exits of GRA. Vehicle attributes have been tuned so that the average speed values measured per lane and per segment from the mobility simulation software match with the corresponding values estimated by the inferred experimental data.

5 Performance Evaluation

The map of the GRA was accurately imported from Openstreetmap, including all the information about speed limits and lanes.

Vehicular mobility simulation has been carried out by using the micro-traffic simulator SUMO (Simulation of Urban Mobility, v0.15.0 [Beh11]) fed with the OD matrix derived as in Section 4. It is a microscopic traffic simulator, able to generate a micro-mobility pattern that captures real drivers behavior. SUMO implements a car following model, accounts for speed limits, lane changing and vehicle overtaking. Given vehicle space-time trajectories produced

with SUMO, the simulation of V2V communications has been implemented in NS-2 [TNS11].

The main simulation parameters are summarized in Table 1. For every scenario, the vehicular simulation lasts 1 hour, and statistics are collected from the simulation trace only in the last 300 s, so as to let the transient for vehicle distribution die out. The monitoring protocol samples vehicle according to the hop length, that in turn depends on the communication link. For a correct reception of a MAC frame, it must be SNR $> \gamma$, where γ is a SNR threshold depending on modulation and coding of the transmitted signal. With our parameter setting (QPSK with coding rate 1/2, at 6 Mbps), it is $\gamma = 8\ dB$. We have SNR $= G(d)P_{tx}/P_N$, where $G(d)$ is the path gain at distance d (two ray ground model on NS2), P_{tx} is the transmitted power, and P_N is the background noise power, equal to $-104\ dBm$ in our case. By setting P_{tx}, we can obtain different target values of the hop length. So, for each chosen value of P_{tx}, the maximum range R_{max} is found from $G(R_{max})P_{tx} = P_N\gamma$.

Table 1: Simulation parameter values

Parameter	**Value**
Road length (km)	68.2
Number of lanes	3
Average vehicle density (veh/km)	31.02
SUMO Simulation duration (s)	3600
Network Simulation duration (s)	300
Frequency of generated messages (msg/s)	1
Transmission Power (mW)	500, 260, 100, 16, 8.7
R_{max} for each tx power value (m)	830, 700, 550, 350, 300
Max forwarding delay, T_{max} (ms)	100
Link Rate (Mbit/s)	6
MAC, PHY parameters	IEEE 802.11p
Propagation Model	Two ray ground

The vehicular traffic monitoring process as described in Section 3 is started by a RSU located in the S-E quadrant of the GRA (exit 19) and by multi-hopping the message returns to the same RSU, after touring the whole GRA in a time ranging between about 0.3 s for a hop length of 830 m up to about 2.6 s for 300 m, on the average.

The average speed values estimated from the data collected by the VANET based monitoring protocol are listed in Table 2 for each of the 29 segments of the GRA and as an overall average. The relative error is defined as $E = |\overline{V}_{VANET} - \overline{V}_{SUMO}|/\overline{V}_{VANET}$. The average number of vehicles sampled for the considered values of P_{tx} and hence of R_{max} are $[82, 101, 129, 202, 242]$, as R_{max} goes from $830\ m$ down to $300\ m$. As the hop length decreases, the probability of disconnections and of failing to complete the GRA tour grows up. Message loss rate grows from about 5% for $R_{max} = 830\ m$ up to 96% for $R_{max} = 300\ m$.

In spite of the quite small number of sampled vehicles (in the order of one or a few hundreds out of several thousands) and of the possibly large message loss rate, the average speed estimate obtained by the monitoring protocol attain a good approximation of the "true" value taken directly from SUMO. Errors of per segment estimates are occasionally between

Table 2: Speed in km/h when sampling distance varies

Seg#	SUMO	830m	E (%)	700m	E (%)	550m	E (%)	350m	E (%)	300m	E (%)
1	79,88	85,50	6,57	83,46	4,28	89,77	11,01	79,59	0,37	82,86	3,59
2	95,23	98,12	2,94	97,53	2,35	96,85	1,67	98,11	2,93	96,68	1,50
3	92,41	93,93	1,63	93,65	1,33	93,99	1,69	90,98	1,57	89,25	3,54
4	90,61	98,47	7,98	100,04	9,42	97,86	7,40	98,14	7,67	92,53	2,07
5	87,41	91,30	4,26	89,61	2,46	96,29	9,22	93,62	6,64	91,60	4,58
6	76,47	74,42	2,76	80,79	5,34	81,60	6,28	79,75	4,10	82,26	7,03
7	79,52	89,18	10,83	90,03	11,67	90,09	11,73	90,39	12,02	86,02	7,55
8	86,35	98,32	12,18	99,27	13,02	100,19	13,82	99,90	13,56	102,39	15,67
9	87,74	95,25	7,88	96,06	8,66	95,30	7,93	98,02	10,49	98,69	11,10
10	88,31	96,66	8,64	93,38	5,43	96,43	8,42	91,03	3,00	86,82	1,71
11	88,24	85,95	2,66	86,51	2,00	83,26	5,97	89,97	1,93	83,48	5,69
12	86,43	90,63	4,63	88,35	2,17	86,11	0,38	86,03	0,47	92,01	6,06
13	88,42	90,59	2,39	91,63	3,50	90,74	2,56	90,20	1,97	83,95	5,32
14	90,43	97,15	6,91	92,64	2,39	97,83	7,56	99,36	8,99	100,82	10,30
15	90,43	79,80	13,32	70,79	27,74	81,49	10,97	89,15	1,44	95,38	5,19
16	90,43	89,78	0,73	90,76	0,37	91,05	0,67	90,56	0,14	94,32	4,13
17	90,40	82,17	10,02	91,22	0,90	86,31	4,74	89,73	0,75	93,19	2,99
18	89,00	88,39	0,69	90,95	2,14	91,41	2,63	92,83	4,12	88,30	0,80
19	85,58	92,84	7,82	90,88	5,83	90,38	5,31	86,82	1,43	85,78	0,23
20	88,77	92,61	4,14	91,60	3,08	95,92	7,45	89,11	0,38	88,56	0,24
21	89,21	81,96	8,85	87,38	2,09	87,24	2,26	91,08	2,05	97,89	8,86
22	89,21	92,30	3,35	92,86	3,94	93,38	4,47	95,01	6,10	96,02	7,10
23	88,35	94,60	6,61	94,69	6,70	83,84	5,38	91,22	3,15	97,76	9,63
24	87,95	95,06	7,48	94,58	7,01	94,24	6,67	92,65	5,07	94,89	7,31
25	91,10	93,85	2,93	95,15	4,26	96,75	5,84	94,52	3,61	97,78	6,83
26	89,93	96,01	6,33	92,33	2,60	89,39	0,60	92,13	2,39	82,57	8,91
27	91,47	95,96	4,68	94,85	3,56	97,77	6,44	98,42	7,06	100,85	9,30
28	91,48	93,19	1,84	92,99	1,63	93,30	1,96	92,16	0,75	90,88	0,65
29	87,40	85,61	2,10	89,66	2,51	86,81	0,68	88,00	0,68	87,38	0,02
Avg	88,21	91,02	5,63	91,16	5,12	91,57	5,58	91,67	3,96	91,76	5,45

10% and 15% and in one case it reaches 27.7%, due to sampled vehicles that are slower or faster than average. In any case, this still provides an information which is considerably good when trying to estimate traffic or problems on road segments.

Result accuracy is only marginally affected as we decrease R_{max}, until we get under 500 m. At 350 m we notice an improvement in the average speed estimation and we also notice a substantial decrease of the peaks that we experience with other samplings. If we try to sample vehicles with shorter distances with respect to 350 m (e.g., 300 m) we see that the error gets worse. This happens because of too many disconnections that do not allow the effective message dissemination. Hence, there is an optimum sampling hop length for the communication protocol.

These results demonstrate that by sampling a little percentage of the vehicles with a message rate that only puts a very light load on the VANET (1 message per second), the vehicles speed estimation can be accurate and above all it is obtained in *real time* and can be refreshed very often (e.g., every 5 min) with very low cost (no infrastructure is required,

except a single RSU for a 68 km road span).

6 Conclusions And Further Developments

The paper has introduced a new communication protocol for VANET networks aimed at detecting road traffic conditions. A simulation experiment has been conducted in a realistic traffic context by applying SUMO and N2 software to simulate the road traffic and the communication network, respectively. The test case consists of a 68 km long ring road expressway around the town of Rome, simulated in the peak hour traffic conditions.

The results of the experiment have shown that the designed communication protocol can perform a monitoring of the whole expressway every 2.6 s or at a higher rate, depending on the sampling rate. The average error of speed estimation ranges from 3% to 5%, as for the whole road; the maximum speed estimation error on a single road segment is about 27%. Such performances are very suitable for a traffic monitoring system addressed to incident detection and real-time traveler information systems. Since traffic phenomena are affected by a large randomness, a very high accuracy of the instantaneous speed estimation is not necessary for these applications. However, a very quick update is crucial to detect traffic instabilities promptly. Finally, the VANET traffic monitoring minimizes the use of cellular telephone network and considerably reduces the high communication costs beard by the vehicular monitoring systems based on GPS positioning and GPRS/UMTS communications.

References

[Ban01] S. BANA and P. VARAIYA: "Space division multiple access (SDMA) for robust ad hoc vehicle communication networks". In: *Proc. of IEEE Intelligent Transportation Systems Conference, 2001*. Oakland, California, 2001, pp. 962–967. DOI: 10.1109/ITSC.2001.948791.

[Beh11] M. BEHRISCH, L. BIEKER, J. ERDMANN, and D. KRAJZEWICZ: "SUMO - Simulation of Urban MObility: An Overview". In: *SIMUL 2011, The Third Int. Conf. on Advances in System Simulation*. Barcelona, Spain, Oct. 2011, pp. 63–68.

[Che12] L. CHENG and S. PANICHPAPIBOON: "Effects of Intervehicle Spacing Distributions on Connectivity of VANET: A Case Study from Measured Highway Traffic". In: *IEEE Communications Magazine* 50 (Dec. 2012), pp. 90–97.

[Far12] H. FARAH, H. KOUTSOPOULOS, M. SAIFUZZAMAN, R. KÖLBL, S. FUCHS, and D. BANKOSEGGER: "Evaluation of the effect of cooperative infrastructure-to-vehicle systems on driver behavior". In: *Transportation Research Part C: Emerging Technologies* 21.1 (2012), pp. 42–56.

[Ger12] M. GERLA, J.-T. WENG, E. GIORDANO, and G. PAU: "Vehicular testbeds - Model validation before large scale deployment". In: *Journal of Communications* 7.6 (2012), pp. 451–457.

[Hua04] Q. HUANG and R. MILLER: "Reliable Wireless Traffic Signal Protocols for Smart Intersections". In: *14th Annual Meeting of ITS America*. 2004, pp. 1–1.

[Kia12] R. KIANFAR, B. AUGUSTO, A. EBADIGHAJARI, U. HAKEEM, J. NILSSON, A. RAZA, R. TABAR, and H. WYMEERSCH: "Design and Experimental Validation of a Cooperative Driving System in the Grand Cooperative Driving Challenge". In: *IEEE Transactions on Intelligent Transportation Systems* 13.3 (2012), pp. 994–1007.

[Leo11] I. LEONTIADIS, G. MARFIA, D. MACK, G. PAU, C. MASCOLO, and M. GERLA: "On the Effectiveness of an Opportunistic Traffic Management System for Vehicular Networks". In: *IEEE Transactions on Intelligent Transportation Systems* 12.4 (2011), pp. 1537–1548. ISSN: 1524-9050. DOI: 10.1109/TITS.2011.2161469.

[Mit10] G. MITROPOULOS, I. KARANASIOU, A. HINSBERGER, F. AGUADO-AGELET, WIEKER, H., H.-J. HILT, S. MAMMAR, and G. NOECKER: "Wireless Local Danger Warning: Cooperative Foresighted Driving Using Intervehicle Communication". In: *IEEE Transactions on Intelligent Transportation Systems* 11.3 (2010), pp. 539–553. ISSN: 1524-9050. DOI: 10.1109/TITS.2009.2034839.

[Mor10] Y. L. MORGAN: "Notes on DSRC & WAVE Standards Suite: Its Architecture, Design, and Characteristics". In: *IEEE Communications Surveys & Tutorials* 12.4 (Oct. 2010), pp. 504–518. ISSN: 1553-877X. DOI: 10.1109/SURV.2010.033010.00024.

[Naf12] N. NAFI and J. KHAN: "A VANET based intelligent road traffic signalling system". In: *Australasian Telecommunication Networks and Applications Conference*. Brisbane, Australia, 2012.

[Nas12] N. NASIRIANI and Y. FALLAH: "Performance and fairness analysis of range control algorithms in cooperative vehicle safety networks at intersections". In: *Conference on Local Computer Networks*. 2012, pp. 848–855.

[Ni99] S.-Y. NI, Y.-C. TSENG, Y.-S. CHEN, and J.-P. SHEU: "The broadcast storm problem in a mobile ad hoc network". In: *Proceedings of the 5th annual ACM/IEEE international conference on Mobile computing and networking*. MobiCom '99. Seattle, Washington, USA: ACM, 1999, pp. 151–162. ISBN: 1-58113-142-9. DOI: 10.1145/313451.313525.

[Pan12] S. PANICHPAPIBOON and W. PATTARA-ATIKOM: "A Review of Information Dissemination Protocols for Vehicular Ad Hoc Networks". In: *IEEE Communications Surveys Tutorials,* 14.3 (Aug. 2012), pp. 784–798. ISSN: 1553-877X. DOI: 10.1109/SURV.2011.070711.00131.

[Qiu12] H. QIU, I.-H. HO, and C. TSE: "A stochastic traffic modeling approach for 802.11p VANET broadcasting performance evaluation". In: *IEEE 23rd Int. Symp. on Personal Indoor and Mobile Radio Communications (PIMRC)*. 2012, pp. 1077–108.

[Sal13] P. SALVO, M. DE FELICE, F. CUOMO, A. BAIOCCHI, and I. RUBIN: "Timer-based distributed dissemination protocols for VANETs and their interaction with MAC layer". In: *77th IEEE Vehicular Technology Conference (VTC2013-Spring)*. Dresden, Germany, June 2013, pp. 1–6.

[Sep13] M. SEPULCRE, J. GOZALVEZ, and J. HERNANDEZ: "Cooperative vehicle-to-vehicle active safety testing under challenging conditions". In: *Transportation Research Part C: Emerging Technologies* 26 (2013), pp. 233–255. ISSN: 0968-090X. URL: http://www.sciencedirect.com/science/article/pii/S0968090X12001258.

[Som11] C. SOMMER, R. GERMAN, and F. DRESSLER: "Bidirectionally coupled network and road traffic simulation for improved IVC analysis". In: *IEEE Transactions on Mobile Computing* 10.1 (2011), pp. 3–15.

[Sun00] M.-T. SUN, W.-C. FENG, T.-H. LAI, K. YAMADA, H. OKADA, and K. FUJIMURA: "GPS-based message broadcast for adaptive inter-vehicle communications". In: *52nd IEEE Vehicular Technology Conference (VTC2000-Fall)*. Vol. 6. 2000, pp. 2685–2692. DOI: 10.1109/VETECF.2000.886811.

[TNS11] *The Network Simulator - ns-2*. 2011. URL: http://nsnam.isi.edu/nsnam/index.php/User_Information.

[Yin13] K. YIN, X. B. WANG, and Y. ZHANG: "Vehicle-to-vehicle connectivity on two parallel roadways with a general headway distribution". In: *Transportation Research Part C: Emerging Technologies* 29 (2013), pp. 84–96.

[Zea12] S. ZEADALLY, R. HUNT, Y.-S. CHEN, A. IRWIN, and A. HASSAN: "Vehicular ad hoc networks (VANETS): status, results, and challenges". English. In: *Telecommunication Systems* 50.4 (2012), pp. 217–241. ISSN: 1018-4864. DOI: 10.1007/s11235-010-9400-5.

Corresponding author: Gaetano Fusco, University of Roma "La Sapienza", Department of Civil, Constructional and Environmental Engineeering, 00184 Roma, Italy, phone: +39 06 44585128, e-mail: gaetano.fusco@uniroma1.it

ITS Solutions for Air Cargo Revenue Management

Tatjana Bolic, Lorenzo Castelli, Desirée Rigonat
Università degli Studi di Trieste

Abstract

In this work we present Revenue Management applications for air cargo and discuss the most relevant issues that currently limit their effectiveness. We explain how these concerns may be tackled by decreasing uncertainty on customer data and improving communication along the supply chain and we illustrate how integration with Intelligent Transportation Systems may play a key role in delivering such improvements.

Keywords: Air transport, Air cargo, Revenue Management

1 Introduction

The International Air Transport Association (IATA) estimated that air cargo is a market that accounts for tens of millions of dollars in revenue for airlines every year (for updated forecasts on market development see https://www.iata.org/publications/economics); both cargo and passenger markets sunk into recession following the 2008 crisis. Passenger market has been increasing since then, while cargo market stabilised and shows some weak signs of reprise in certain markets (http://www.iata.org/pressroom). As the air cargo market is foreseen to grow steadily in future years, many experts view this as a right moment to start a systematic application of Revenue Management (RM) to the air cargo part of business for passenger airlines (see for example [Ama11] and [Mor09]). So far, no more than 35% of the largest passenger airlines have invested in RM for air cargo, according to a research by Air Cargo World [Med12].

Intelligent Transportation Systems (ITS) that we refer to here are Information and Communication Technologies (ICT)-based solutions that aim at improving efficiency from a functional and environmental point of view in most fields of transportation. Making transportation "smarter" can involve either the development of new applications and solutions or the improvement of existing ones. Airline industry is relying on ICT applications for decades already: seat reservation, air traffic management, check-in and so on. Air cargo is no exception, and all the reservation-related decisions are managed through complex ICT systems. There

is, however, much innovation coming from other ITS solutions that could affect positively the airlines' air cargo business, improving the efficiency of the service. Here we will consider only the innovations that affect RM application within air cargo.

To further the discussion, air cargo RM has different requirements from the one for passengers. While on one hand academic research on cargo RM has been extremely productive in the last 10 years (see for example [Ama07][Bil03][Kas96] and [Pop06]), when it comes to actual implementations of commercial software there is very little information on how the RM process is performed. This is obviously due to restrictions and non-disclosure agreements regarding the intellectual property of the software houses. This also proves how crucial RM techniques have become for any cargo booking/management system that wishes to compete on the market; practically speaking, RM has become a part of the core business for these software companies. However, there is still a lot of room for improvement that can be explored through research, as suggested by the aforementioned literature.

The reminder of this paper is organised as follows: Section 2 describes the characteristics of RM and the classes of problems that if resolved could make RM more efficient, Section 3 links the innovations coming from ITS that could better the current RM systems with the particular RM characteristics, and Section 4 offers conclusions and our expectations on the future development of RM and ITS-RM links.

2 The air cargo revenue management (RM) system

An RM system generally consists of four main components (see [Tal05]):

- Data-collection engine: collects and stores historical data for analysis (i.e., prices, demand, causal factors).
- Forecasting and estimation engine: forecasts relevant quantities (demand, capacity, cancelation rates, show-ups) based on historical data.
- Optimisation engine: using the forecasted values, finds the optimal set of controls (capacity allocations, prices, overbooking levels) to be applied until the next re-optimisation.
- Control engine: applies the optimised controls to sell inventory.

The RM process typically iterates through these steps at repeated intervals, the frequency of which varies by industry. In air cargo the re-optimisation is usually run on a daily basis in the four weeks prior to flight departure, less frequently in the preceding weeks (see [Sab04]). This is due to the fact that in the air cargo industry a consistent portion of booking requests usually comes in during the few days before the take-off. Traditionally, before the application of RM, air cargo was mostly managed through a "first come, first served" or a "max loading" approach [Pop06]. Neither of these approaches guarantees optimal choices related to profitability that, instead, can be significantly enhanced by addressing the following five classes of problems, sometimes referred to as "techniques" or "applications" [Sab04]:

Capacity forecasting. Capacity available for future flights can be estimated based on aircraft dimension and configuration, where parameters such as passenger, baggage and fuel weight can be predicted using the flight's historical data. The capacity is also affected by the estimated weight of *air-mail* parcels, the size of allotments (i.e., weight and container dimensions) and eventually, environmental conditions (i.e., weather forecasts).

Demand forecasting and overbooking. Demand forecasting aims at estimating how much cargo will tender for a particular flight. Having the historical and present data on bookings, it is possible to forecast the cargo demand in terms of day of the week of departure etc. One important component of Demand Forecasting is the estimation of cancellations and of no-show rates; these allow for the calculation of an appropriate overbooking limit. Overbooking is a well-known practice among airlines and, in general, service-based companies, who sell more units of capacity than those physically available. For air cargo the objective is to set the overbooking limit in order to maximise space utilisation at departure (and hence, to maximise revenue) and to minimise the spoiled (unutilised) capacity and the cost for off-loading and re-routing of the excess cargo.

Allotment management. Allotments are long term contracts (usually 6 or 12 months) for space reservation on future flights; such agreements are generally stipulated between an airline and a customer, such as a freight forwarder, who generates a big amount of traffic. Allotment management analyses historical behaviour of customers in order to determine the corresponding show-up rate. From this data the airline is able to determine whether a contract should be accepted or not. While most cargo revenues usually come from such few, big customers, granting an allotment to a customer that systematically leaves his space unused increases spoilage cost. Additionally, if anticipated demand for the flight yields higher revenues than the one coming from the contract, allotting that space to the contract results in a revenue loss.

Price optimisation and capacity management. The primary purpose of price optimisation in RM is the estimation of the value created for customers and subsequent price setting to capture that value; it is tightly coupled with capacity management, that instead deals with pricing of the available capacity in order to maximise revenue. As a direct consequence, capacity management establishes the criteria by which a booking for a shipment should be accepted or rejected. Differently from the passenger scenario, capacity management in cargo is a difficult problem to solve because of the three-dimensionality of the shipments. Thus the loading strategy, together with the booking acceptance strategy is crucial to the effectiveness of RM. The decision whether to accept or reject a reservation can be based on different control strategies, the most common in the passenger airline segment are:

- Booking limits: a fixed amount of capacity is assigned to each fare class;
- Protection levels: a certain amount of capacity is reserved exclusively for a class (or a set of classes);

- Bid pricing: a threshold value is set and only requests that generate a revenue that is higher than the threshold are accepted; the threshold is updated every time a booking is accepted.

Note that the first two are capacity based controls while the third is revenue based. In the air cargo industry bid pricing is by far the most commonly used, mostly because it has an intuitive meaning and usually guarantees good revenues [Tal05]. The bid price for a flight represents the minimum amount that should be charged for the cargo, and is proportional to the level of the demand: when demand is low the bid price is low (trying to avoid capacity spoilage), when the price is high the bid price is high (avoiding exceeding the available capacity).

Routing optimisation. Calculates the optimal route for shipments in order to maximise the revenue for the airline. The aspect of network generation (to obtain all possible routes from origin to destination) and the one of network optimisation (to optimally allocate demand to the available capacity) need to be considered. Air cargo routing optimisation can be carried out at leg level, segment level, or origin and destination (O/D) level, depending on the structure of the network and the data that is available on user behaviour. Network optimisation is also necessary in the case of oversale, that is, when a shipment has to be deviated on another route because the assigned flight was overbooked. The new route should respect the schedule for delivery while still being profitable for the company, and take into account the availability of the cargo handling equipment along the route.

3 ITS and air cargo revenue management

The strength (and key to success) of ITS solutions is interoperability and mutual positive influence, meaning that improving one system has beneficial fallouts on several others that are related or operate together with it. As illustrated in the previous section, applying RM to air cargo and enhancing its effectiveness involves addressing several sub-problems. The issues listed below arise in the mentioned sub-problems, and can be responsible for reduction of the effectiveness of RM [Tal05]:

- Uncertainty on the available capacity for cargo;
- Uncertainty of the cargo size due to multiple dimensions and density;
- Loss of revenue coming from no-shows since they are usually not charged;
- Loss of revenue coming from oversales due to the necessity to re-route or cancel a booking.

We can identify three main causes for all the issues above: the first deals with the way contracts are stipulated with long term customers, the second with uncertainty of users'

behaviour, the third with uncertainty of shipment data. Allotment contracts are the factor influencing the most the performance and revenues in air cargo transportation but are also a source of inefficiency where ITS can deliver only minor improvements. In fact, once a contract is settled, there is little an airline can do to guarantee that it will be as profitable as expected [Sab04]. Under contracts today, no-shows are not charged for spoiled space and no incentives are offered to customers to inform the airline when they are not going to use their space that would otherwise be made available for sale to other customers. Changing contracts would require both airlines and forwarders to disclose information that is usually kept confidential. [Ama07] and [Ama11] study the role of asymmetrical information on customers' demand, operating cost, margin and reservation profit. They investigate under which conditions the maximum combined profit of the involved agents can be obtained in the presence of such an asymmetry and propose new contract models that take this into account. Uncertainty on users' behaviour and shipment data, on the other hand, can be tackled by providing a continuous, reliable and detailed information flow. Giving the RM system access to data originating in other ITS applications, i.e., electronic shipment documentation, or tracking data, may improve the effectiveness of RM, directly and indirectly. New and more reliable information would have direct impact on RM by enabling development of innovative optimisation models. Furthermore, improvement of communication among the supply chain actors and of information flow along the whole cargo transportation process could have significant, if indirect impact on RM.

3.1 Improving demand forecasting

Commercial software packages today typically describe demand through a stochastic process which takes into account several factors, such as historical data, seasonality and expected weather conditions. Demand is forecasted by clustering previous booking data into classes based on revenue and density of the shipment [Sla04]. This classification enables forecasting and optimisation to be performed by rate and load mix: for each of these classes, the system estimates the volume of each revenue type to be tendered at departure. Moreover, RM systems are usually able to predict demand at leg, segment or O/D level and to optimise accordingly.

The underlying mathematical models however are far from being ideal for the cargo scenario. In fact, traditionally, demand forecasting for air cargo has been carried out by adapting models that were first developed for passenger RM. This approach is far from being optimal, since cargo shows no or very little of the recurrence patterns that are instead easy to recognise for passengers (i.e., more traffic on certain routes during holidays). The factors that can affect cargo demand are usually erratic and difficult to forecast, i.e., harvests, climatic events, trends and special events, and so on. A new approach to RM, called Customer Centric Revenue Management, has recently been developed within the hospitality industry, and is currently spreading among passenger transportation businesses, especially railways and airlines. In [Vin08] Customer-Centric Revenue Management is defined as a Customer-Relationship

Management (CRM) enabler to increase an airline's profitability based on customer insight. It requires a combination of marketing, revenue management and real-time inventory control to facilitate one-to-one targeted responses to manage the customer's life-cycle. Real-time inventory control could be provided through integration with freight tracking ITS solutions, paired with electronic documentation, allowing the determination of the freight status in real-time (i.e., did it obtain customs clearance). Applying this novel approach implies developing new demand forecast models that use the historical and real-time data and thus can based on the customer behaviour analysis as well.

According to [Fre07] benefits coming from the adoption of an RM system that is customer-centric include lowering customer costs, providing shipping alternatives, and tracking network performance. More recent sources, such as [Mor09] consider customer-centricity among the best practices for a successful RM strategy in the cargo field.

3.2 Improving communication across supply chain

Unfortunately, a lot of time is wasted in the air cargo supply chain due to poor coordination and communication between shippers, freight forwarders, airlines and the other actors [Pop06]. This is related to RM as long as providing more precise data decreases the uncertainty that must be taken into account when applying RM optimisations. At the same time, inefficiencies and errors in communication affect the effectiveness of RM: optimisations become useless if the shipments are not cleared from customs on time for the scheduled take-off for example. The only viable way to solve these issues lies in the adoption of common protocols in communications, to allow interoperability among different information systems. In general, when aiming at interoperability, several levels of standardisation can be applied, i.e., on the data format, on the message structure, on the communication protocol or on the whole information systems (i.e., by imposing adoption of the same software package). We believe that in order to obtain a smooth flow of information across the supply chain actors, standardisation of data format and message structure would be enough, since the Internet delivers a well established set of communication protocols (on which all ITS already rely) and system-level standardisation would deliver no further benefits in this context.

Today, one major reason for inefficiency within cargo industry is related to shipment documentation: a shipment can be accompanied by up to 30 documents, ranging from invoices over airwaybills to customs declarations [PTV09]. Handling of these documents in paper form is highly inefficient: it requires a lot of time and manual check is often needed, causing further delays. To face these issues, solutions like Intelligent Cargo (for generic freight transportation, see http://www.euridice-project.eu and http://i-cargo.eu for further reference) and e-freight (specific to air traffic, see http://www.iata.org/whatwedo/cargo/efreigh) have been proposed. As stated by IATA, "e-freight aims to take the paper out of the air cargo supply chain and replace it with cheaper and more reliable electronic messaging". E-freight is an initiative involving shippers, carriers, freight forwarders, ground handlers, and customs authorities that uses the existing air cargo industry messaging infrastructure and relies on

open standards such as EDI (Cargo-IMP or XML) or common electronic image file formats (for scanned documents).

Electronic data on shipments can be even more beneficial to the industry if it can be exchanged through a common infrastructure. Today, most providers for cargo RM and reservation software are migrating their products to a cloud-computing infrastructure, more specifically to the SAAS (Software As A Service) paradigm. For many businesses, cloud computing is a cost-savvy option when compared to having an in-house dedicated IT department. It offers more flexibility, since: resources can be rented on a usage basis; device and location independence; resources are accessible from anywhere over the internet and are more secure; bug fixes and updates are installed by the service provider instead by the end user. This innovation is relevant, since it allows access to RM systems by potentially every airline, including those that are too small to afford the purchase of dedicated computer hardware. Thanks to Cloud Computing, these airlines can rent the computing power from the RM software vendor and can access the application through web browsers.

From the customer's perspective, a combination of user-tailored demand prediction (as described in sec. 3.1) and standardised formats for data exchange could lead to the development of highly customisable products (i.e., optional services that are selectable rather than bundled together) and aggregation solutions (i.e., rate comparison services). Certain operators already offer services such as alerts, tracking, special offers, pick-up reservation, delivery notification to their customers through emails, websites or mobile applications. These are so far isolated solutions that are not interoperable and do not offer any data interchange services.

3.3 Information on cargo and ground operations

Ground operations are not strictly connected to RM, since they deal with cargo manipulation once it is delivered to the airline's warehouse. However, inefficiency in ground operations can make RM optimisations unreliable, when, for example, a shipment expected for a certain flight did not arrive on time to the warehouse, or the shipment was tampered with, or it is not cleared at customs, and so on. The efficiency of ground operations affect not only the RM effectiveness but also that of the whole transportation service, so the same considerations made for communication among the supply chain actors apply here as well. The paradigm of Intelligent Cargo (see [PTV09]) states that the wide adoption of advanced systems for freight localisation and identification is likely to make hub and transport operations faster and smoother; this holds true for ground operations at the airports too. ITS can help automate some and speed up most ground operations, thus making them smoother and reducing the need for manual checks, which is usually greatly expensive in terms of time.

Ground operations and information systems for ground operations that could benefit from ITS usage are the following [PTV09]:

- Border clearance: in order for freight to be cleared, its transportation documents must be checked. Electronic documentation (as previously discussed) circumvents manual

control of up to 30 paper documents per cargo shipment.

- Remote freight information (RFI): it is of vital importance to know the characteristics of freight (i.e., hazardous material, perishable items etc.) in order to know how to handle it properly. RFI is a major ITS application in road transport and is relevant to air transport, especially in an inter-modality context.

- Weight screening: precise data on cargo weight can improve RM decisions. Moreover, it can speed up border and customs clearance.

- Warehousing: efficient storage and localisation influence time required for loading operations. Indoor localisation and identification enable better management of warehouses.

To this end, the positioning (i.e., GPS) and identification technologies (i.e., RFID) combined with wireless communication technologies (i.e., mobile networks) are key enablers. Some pilot projects aim to integrate all these technologies on the cargo boxes, the so-called "smart Unit Load Devices (ULDs)" that should provide the following [Pan07] :

- Theft of ULDs will become difficult, if not impossible;

- Accuracy of current location data of ULD will result in savings when airlines know for certain where the asset was last time recorded, at which time, and in which agency's control area;

- Make other airlines/Ground Handling Agencies accountable while taking custody of airlines' ULDs during interline transfers;

- Faster turnover of assets over network, thus increasing the utilization rate, and reducing ULD inventories;

- Standardisation of cargo handling processes at the airports.

4 Expectations and conclusions

In the present work we identified a roadmap towards growing automation and integration that could effectively solve some issues that are currently weighting on the cargo supply chain, starting from inter-modal operations and warehousing, up to travel documents such as airwaybills and customs clearance papers. Clearly, some of these expectations are easier to fulfil than others. On one hand, for example, IATA estimates that e-freight will be adopted by 100% of airports within 3 or 4 years. On the other hand, most technologies employed for warehousing automation (i.e., RFID tags) do lack a common operational standard, and are thus lagging behind where interoperability is concerned. We thus expect that a longer time is required for implementation. As soon as common standards for interoperability are defined, however, we expect technology adoption to run smoothly.

From the RM point of view, implementing a higher level of automation in ground operations and related information services implies that real-time information on freight status will become available. If it is also made accessible in a standard data format, then it can be used for making capacity management dynamic, thus reducing capacity spoilage and delivering a more efficient service. Exploiting the wide availability of real-time data generated from other ITS is probably the fulcrum of future development in RM. We can expect several improvements coming from a "real-time enabled" Revenue Management approach: a reduction of spoiled space due to improved capacity management, better quality of service due to more reliable predictions, more services and options available to customers due to increased flexibility.

From the algorithmic point of view, since RM for air cargo has received growing interest as a research topic in the last years, we expect innovative and improved models to be developed in the near future, and implemented in new generation of software products shortly after. We expect user behaviour prediction to be the first to be tackled due to the growing popularity of machine learning and data mining techniques, paired with the massive amounts of data (either user or machine generated) that shall be provided through integration with ITS. New dynamic capacity management algorithms that can take advantage of real-time data are expected to follow shortly after. Innovations in routing will most likely involve multi-leg flights and optimisation spanning over different modes of transportation.

Most of the software technologies that are helpful for RM deployment in air cargo, are already mature or in an advanced stage of development. Cloud computing is a good example of this; we faced a massive migration of software and services to a cloud-based environment in recent years. We can thus only expect this trend to continue and, as far as RM is concerned, we can also assume that software solutions are going to completely migrate to the cloud paradigm within the next couple of years. In order to fulfil this, the only underlying necessity we can identify is the wide availability of high-speed Internet connections.

From the hardware based ITS point of view (i.e., smart ULDs, freight tracking, freight identification) most applications have more than one alternative for technological implementation. For example identification nowadays can use either barcodes or RFID tags. It is not clear at the moment which level of standardisation each of these applications will require, but in order to allow interoperability among systems some common standard is required at least at data and message format level. The Internet of Things relies on open standards such as XML for sharing sensor data among smart appliances (i.e., for home automation). A similar approach is likely to be a viable choice for sharing data among ITS solutions as well.

To sum-up, many ITS applications and hardware that RM can benefit from are in the range from widely available to close to being available. In order to enable improvement in air cargo RM these need to be in wide-spread use and offer interoperability within the supply chain. The improvements to be gained are two-fold: more sources and real-time data that offer a base for improved models within RM, and at the same time inform and keep up-to-date the RM system, thus improving its efficiency.

References

[Ama07] A. AMARUCHKUL, W. COOPER, and D. GUPTA: "Single-Leg Air-Cargo Revenue Management". In: *Transportation Science* 41.4 (2007), pp. 457–469.

[Ama11] A. AMARUCHKUL, W. COOPER, and D. GUPTA: "A Note on Air-Cargo Capacity Contract". In: *Production and Operations Management* 20.1 (2011), pp. 152–162.

[Bil03] J. BILLINGS, A. DIENER, and B. YUEN: "Cargo revenue optimisation". In: *Journal of Revenue & Pricing Management* 2.1 (2003), pp. 69–79.

[Fre07] L. C. FREELAND: "Adoption of customer-centric cargo revenue management: Brief history of cargo revenue management vs passenger revenue management". In: *Journal of Revenue & Pricing Management* 6.4 (2007), pp. 284–286.

[Kas96] R. KASILINGAM: "Air cargo revenue management: Characteristics and complexities". In: *European Journal of Operational Research* 96.1 (1996), pp. 36–44.

[Med12] A. MEDEPALLI and J. SOFTWARE: *Entering the Next Dimension of Cargo Revenue Management*. 2012. URL: `http://www.slideshare.net/JDASoftware/entering-the-next-dimension-of-cargo-revenue-management`.

[Mor09] G. MORELLO: *Revenue management and air cargo*. Ed. by I. YEOMAN. London, 2009.

[Pan07] P. N. PANDIT: *Applications of RFID in air cargo*. IINFOSIS White Paper. 2007.

[Pop06] A. POPESCU: "Air cargo Revenue and Capacity Management". PhD thesis. Georgia Institute of Technology, 2006.

[PTV09] PTV PLANUNG TRANSPORT VERKEHR AG AND ECORYS NEDERLAND BV: *Final report Intelligent Cargo Systems study (ICSS) Impact assessment study on the introduction of intelligent cargo systems in transport logistics industry*. 2009.

[Sab04] SABRE INC: *Cargo Revenue Management White Paper*. 2004.

[Sla04] B. SLAGER and L. KAPTEIJNS: "Implementation of cargo revenue management at KLM". In: *Journal of Revenue & Pricing Management* 3.1 (2004), pp. 80–90.

[Tal05] K. TALLURI and G. VAN RYZIN: *The Theory and Practice of Revenue Management*. Springer, 2005.

[Vin08] B. VINOD: "The continuing evolution: Customer-centric revenue management". In: *Journal of Revenue & Pricing Management* 7.1 (2008), pp. 27–39.

Corresponding author: Lorenzo Castelli, Università degli Studi di Trieste, Dipartimento di Ingegneria e Architettura, 34127 Trieste, Italy, phone: +39 040 558 3416, e-mail: castelli@units.it

Air Traffic Optimization Models for Aircraft Delay and Travel Time Minimization in Terminal Control Areas

Marcella Samà, Paolo D'Ariano, Andrea D'Ariano, Dario Pacciarelli
Università degli Studi Roma Tre

Abstract

This work addresses the real-time optimization of take-off and landing operations at a busy Terminal Control Area (TCA) in case of traffic congestion. These areas are becoming the bottleneck of the entire air traffic control system, in particular in the major European airports where there is a limited possibility to build new infrastructure. The problem of effectively optimizing TCA operations is particularly challenging, since a detailed solution, incorporating safety rules, has to be quickly computed. Also, key performance indicators should be considered in order to evaluate the quality of the proposed action plan. This paper proposes new aircraft scheduling and routing models, formulating detailed TCA safety rules and different objective functions. The minimization of the largest delay and the total travel time spent in the TCA are investigated. Computational experiments are performed via a commercial solver on a real test case for Roma Fiumicino airport, the largest Italian airport. Disturbances are generated by simulating various sets of random landing/take-off aircraft delays. This analysis makes possible the selection of those solutions offering the best compromise among the different objectives.

Keywords: Efficient Landing and Take-Off Operations, Microscopic Air Traffic Control Models, Mixed Integer Programming.

1 Introduction

The ever growing demand of air transport is increasing the pressure on air traffic controllers, since air traffic in peak hours is getting closer to the capacity of the Terminal Control Area (TCA). In fact, at least in the major European airports, there is limited possibility of creating new infrastructure. Aviation authorities are thus seeking intelligent methods to better use the available infrastructure and to improve the overall system performance [And13; Cas11;

DAr12; Kim11; Pel12]. However, the development and implementation of effective optimization methods for such operational problems require paying attention to a number of aspects that are rarely taken into account simultaneously in scheduling theory:

- The optimization model should be able to incorporate all detailed information that is important for ensuring the schedule to be compliant with TCA safety regulations, including the capacity of air segments and runways in the TCA. Due to the high level of detail required in the formulation, those information are often neglected in macroscopic models for large networks with multiple airports [Ber11; Chu10; Kuc00].

- The time available for developing a new schedule can be very limited, since a computerized scheduler should be able to promptly react to changes of aircraft position and speed occurring during operations.

- To a large extent, ATC operations and related issues are still scheduled by human controllers, who develop feasible schedules based on their past experience and intuition. Without using any formally defined procedure, there is a lack of a generally recognized performance indicator.

This paper addresses the three items above. The first item is approached by developing new scheduling and routing models for Air Traffic Flow Management in a Terminal Control Area (ATFM-TCA). We started from the approach of Bianco et al. [Bia06] that is based on the no-wait job shop scheduling problem with aircraft routing and timing variables. However, we use the alternative graph model of [Mas02] to increase the level of detail of the ATFM-TCA formulation. Other relevant TCA aspects are modelled, such as holding circles, speed intervals for aircraft, capacitated use of air segments and no-store constraints at runways. Compared with previous works from our research group [DAr10; DAr12; Sam13b; Sam13a], we optimize routing and scheduling simultaneously.

The second and third items require to test the optimization model with a short computation time and to pay special attention to the definition of the indices. We follow a most common choice in the literature that is the use of a single objective function, typically as a combination of various performance indicators (see, e.g., the reviews in [Bar12; Ben11; Cla10; Koh07]). In fact, single objective approaches are faster than multi-objective optimization.

Computational experiments have been carried out via the commercial solver IBM ILOG CPLEX MIP 12.0. The test bed is the main Italian airport, Roma Fiumicino (FCO). We consider practical size instances with several delayed aircraft and solve the related ATFM-TCA problems. As for the indices, we analyze the minimization of aircraft delays and/or travel times. A weighted sum of them is also investigated in order to identify a reasonable balance.

The structure of the paper is the following. Section 2 formally describes the ATFM-TCA problem, including a description of the specific constraints and objectives, while Section 3 presents the mathematical formulations. Section 4 reports a campaign of experiments. Section 5 summarizes the paper results and outlines on-going research directions.

2 Problem description

The problem of managing aircraft in a TCA can be divided into: (i) Routing decisions, where an origin-destination route for each aircraft has to be chosen; (ii) Timing decisions, where routes are fixed under traffic regulation constraints and aircraft passing timing have to be determined in each air segment, runway and (possibly) holding circle. In this work, routing (i) and timing (ii) decisions are taken simultaneously.

The ATFM-TCA problem is defined as follows: given a set of landing and take-off aircraft and, for each aircraft, a set of possible paths in the TCA, its current position, its scheduled runway occupation time and a time window to accomplish the landing/departing procedures, assign an entry time to each aircraft in each resource (holding circles, air segments, runways) traversed by the aircraft in the chosen path in such a way that the resulting schedule is conflict-free. A potential conflict occurs whenever two aircraft traversing the same resource do not respect the minimum longitudinal/diagonal safety distance.

Figure 1 presents the scheme of Fiumicino Terminal Control Area (FCO TCA). Three runways (RWY 16L, RWY 16R, RWY25) can be used for departing and landing procedures, but two of them (RWY 16R and RWY 25) cannot be used at the same time and are thus considered as one. The airport resources are 3 airborne holding circles (CIA, CMP, TAQ, 1-3), 7 air segments for landing procedures (4-10) and 3 for departing ones (14-16), 2 runways (12-13) and 1 common glide path (11). Black Triangles represent merging points between air segments.

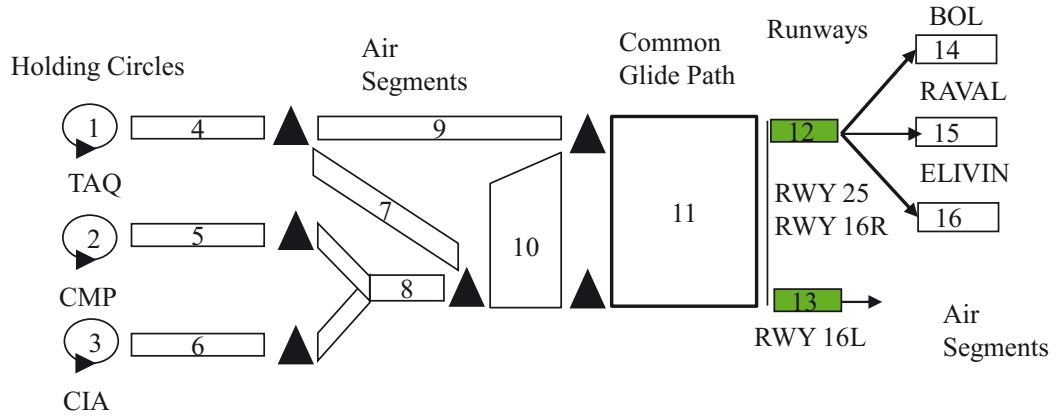


Figure 1: Fiumicino Terminal Control Area

In the TCA, landing aircraft moves from an entry point to a runway, following a standard descend profile, while maintaining a minimum safety distance with the other aircraft, depending on their type and position. Similarly, departing aircraft leaves the runway flying toward the assigned exit point along an ascent standard profile, still respecting separation safety distances. Since the variability of aircraft speed in the TCA is limited, this distance can be translated into a *setup time*. Setup times are sequence-dependent, since the minimum distance between heavy, medium or light aircraft depends on the relative order of processing

of the common resources. Each aircraft has a minimum entry time in the TCA, named *release time*.

The runway can be occupied by only one aircraft at a time, and each aircraft has a *processing time* on a runway and on the air segments before or after it, according to its landing/take-off profile. On the air segments, the processing time can vary within a pre-defined time window, due to the limited possibility of aircraft speed changes.

Once an aircraft enters the TCA, it should proceed to the runway. However, in case of congestion, airborne *holding circles* can be used as buffers until aircraft can be guided through their landing procedure. Once entered a holding circle, the landing aircraft must fly at a fixed speed for a number of half circles, as prescribed by the air traffic controller. We assume that holding circles are uncapacitated and there is a maximum number of allowed half circles.

Departing aircraft instead can be delayed in entering the TCA at ground level, before entering the runway. A departing aircraft is supposed to take-off within its assigned time window and is late whenever it is not able to accomplish the departing procedure within its assigned time window. Following the procedure commonly adopted by air traffic controllers, we consider a time window for take-off between 5 minutes before and 10 minutes after the *Scheduled Take-off Time* (STT). A departing aircraft is considered late if leaving the runway after 10 minutes from its STT. So a ground delay does not necessarily cause a delay at the runway. Arriving aircraft are late if landing after their *Scheduled Landing Time* (SLT).

We use the following notation for the aircraft delays. *Entrance delay* is the delay of each aircraft at the entrance in the TCA. *Exit delay* is the delay of each aircraft at the runway. The latter value is partly a consequence of a possible late entrance, which causes an *unavoidable delay* at the runway, and partly due to additional delays caused by the resolution of potential aircraft route conflicts in the TCA, which is named *consecutive delay* [DAr07; DAr10; DAr12]. A landing aircraft can have a consecutive delay at the entrance, if it is delayed in entering the TCA due to other aircraft scheduled on its entrance air segment. Landing and take-off aircraft can have a consecutive delay on a runway, if they have to give precedence to other aircraft in one or more TCA resources.

Two objective functions are considered: the minimization of the maximum consecutive delay and of the total travel time spent by aircraft in the TCA. The latter can be considered as a good surrogate for the energy consumption.

3 Formulation of the aircraft scheduling and routing problem

In the general job shop scheduling formulation of the ATFM-TCA problem, an *operation* denotes the traversing of a resource (i.e. air segment, common glide path, runway, holding circle) by an aircraft (i.e. job). The sequence of operations of an aircraft represents its *route*. Once a route has been assigned to each aircraft (routing problem), the ATFM-TCA problem (with fixed routes) is reduced to the Aircraft Scheduling Problem (ASP). The variables of the ASP are the start time t_i of each operation i to be performed by an aircraft on a specific resource. For a given operation i, we denote with $\sigma(i)$ the operation following i on its route

and with $w_{i\sigma(i)}$ a minimum processing time of i. A set of feasible route timings is *conflict-free* if, for each pair of operations associated to the same resource, the minimum separation constraints are satisfied.

The ASP can be represented by an alternative graph [DAr10] as follows. Let $G = (N, F, A)$ be the graph composed by the following sets: $N = \{0, 1, ..., n\}$ is the set of *nodes*, where 0 and n represent the start and the end operations of the schedule, while the other nodes are related to the start of the other operations; F is the set of *fixed arcs* that model the sequence of operations to be executed by an aircraft; A is the set of *alternative pairs* that model the sequencing decision. Each pair $((i, j), (h, k))$ is composed by two alternative arcs, either (i, j) or (h, k) must be selected in a feasible schedule.

Each node of the graph is thus associated to the start time t_i of operation i. By definition, the start time of the schedule is $t_0 = 0$. Each arc (i, j), either fixed or alternative, has a length w_{ij}, which indicates a minimum separation time between operations i and j, i.e. $t_j \geq t_i + w_{ij}$. A detailed description of the constraints related to the specific TCA resources can be found in [DAr12; Sam13b; Sam13a].

A *selection* S is a set of alternative arcs, at most one from each pair. A solution is a *complete* selection S, where an arc for each alternative pair of A is selected and it is *feasible* if the connected graph *(N, F, S)* has no positive length cycles. Given a feasible schedule S, a timing t_i for operation i is the length of a longest path from 0 to i ($l^S(0, i)$).

The alternative graph with flexible routing can be viewed as a particular *disjunctive program*. Adding the constraint for the different routes for each aircraft, we let X be the set of feasible ATFM solutions:

$$X = \left\{ \begin{array}{ll} t \geq 0, x \in \{0,1\}^{|A|}, y \in \{0,1\}^{|C|} : & \\ t_{\sigma(i)} - t_i + M(1 - y_{ab}) \geq w_{i\sigma(i)} & \forall (i, \sigma(i)) \in F \\ t_j - t_i + M(1 - x_{ijhk}) + M(1 - y_{ab}) + M(1 - y_{cd}) \geq w_{ij} & \forall ((i, j), (h, k)) \in A \\ t_k - t_h + M x_{ijhk} + M(1 - y_{ab}) + M(1 - y_{cd}) \geq w_{hk} & \\ \sum_{a=1}^{R} y_{ab} = 1 & b = 1, ..., Z \end{array} \right\} \quad (1)$$

The variables are the following: $|N|$ real variables t_i associated to the start time of each operation $i \in N$, $|A|$ binary variables x_{ijhk} associated to each alternative pair $((i, j), (h, k)) \in A$, and $|C|$ real variables associated to the routes of the set of aircraft considered. Further, Z is the number of aircraft, and R the number of routes for each aircraft. The constant M is a sufficiently large number, e.g. the sum of all arc lengths.

Variable $y_{ab}(y_{cd}) \in \{0, 1\}$ indicates if route a (c) is chosen (1) or not (0) for aircraft b (d). Exactly one route for each aircraft must be selected, e.g. for aircraft b: $\sum_{a=1}^{R} y_{ab} = 1$. When a route a is chosen for aircraft b, each constraint related to the fixed arcs of route a and aircraft b must be satisfied, i.e. $t_{\sigma(i)} - t_i \geq w_{i\sigma(i)}$ must hold.

If $y_{ab} = y_{cd} = 1$ and aircraft b and d are scheduled on a same resource of the TCA, a potential conflict exists on that resource and an ordering decision has to be taken. This is modelled by using the variable $x_{ijhk} \in \{0, 1\}$ for the alternative pair $((i, j), (h, k))$, related to

the two aircraft travelling on that specific resource. If $x_{ijhk} = 1$ then $t_j - t_i \geq w_{ij}$ must be satisfied (i.e. $(i, j) \in S$), otherwise $t_k - t_h \geq w_{hk}$ must be satisfied (i.e. $(h, k) \in S$).

In order to formulate the objective functions, we must introduce two types of due date arcs: *entrance due date arcs*, associated with the first operation of each landing aircraft to measure its entrance delay; *exit due date arcs*, associated with the runway operation of each aircraft to measure the consecutive delay in the TCA.

We let j be the first operation of a landing aircraft in the TCA, β_j be its scheduled entrance time and f_j be the entrance delay, that we assume not controllable by the traffic controller. The length of entrance due date arc is $d_j = -\beta_j - f_j$. The consecutive delay at the entrance of the TCA is $max\{0, t_j + d_j\}$.

We let i be the arrival/departure operation at/from a runway r of a landing/take-off aircraft A, β_i be its scheduled arrival/departure time and τ_i be the earliest possible entrance time in the runway r. The total exit delay of A at r is $t_i - \beta_i$.

The total exit delay is composed by the *unavoidable delay* (which cannot be recovered by aircraft rescheduling) plus the *consecutive delay* (required to solve potential conflicts). In the graph, these delays are computed at the runway as follows. Let's fix the exit due date arc length as $d_i = -max\{\tau_i, \beta_i\}$, the unavoidable delay is $max\{0, \tau_i - \beta_i\}$, while the consecutive delay is $max\{0, t_i + d_i\}$.

The Maximum Tardiness **MT** corresponds to the minimization of the maximum consecutive delay [DAr07; DAr10]. Both for the entrance and exit due dates, MT is the largest positive deviation from the entrance and due date times. A feasible schedule S is optimal if $l^S(0, n)$ is minimum over all the solutions. The **MT** formulation is therefore:

$$\begin{aligned} &\min t_n \\ &s.t \\ &t_n - t_k + M(1 - y_{ab}) \geq d_k \quad \forall (k, n) \in F \\ &\{x, y, t\} \in X \end{aligned} \tag{2}$$

The minimization of the Total Travel Time Spent **TTTS** is the objective function we use as a surrogate for the energy consumption in the TCA. For an aircraft a, let t_{af} be the finish time of its last operation and let t_{ar} be its release time, the travel time spent in the TCA by this aircraft can be computed as $t_{af} - t_{ar}$. The **TTTS** formulation is the following:

$$\begin{aligned} &\min \textstyle\sum_{a=1}^{|Z|} t_{af} - t_{ar} \\ &s.t \\ &\{x, y, t\} \in X \end{aligned} \tag{3}$$

The two objective functions reported in (2)–(3) support specific aspects of the ATFM-TCA problem. We also study a convex combination of the two objective functions, named **MT-TTTS**:

$$\begin{array}{l} \min \alpha\delta t_n + (1-\alpha)\gamma\sum_{a=1}^{|Z|} t_{af} - t_{ar} \\ s.t. \\ t_n - t_k + M(1-y_{ab}) \geq d_k \qquad\qquad \forall (k,n) \in F \\ \{x,y,t\} \in X \end{array} \tag{4}$$

where $\delta = 1/t_n^*$, $\gamma = 1/\sum_{a=1}^{|Z|} t_{af}^* - t_{ar}^*$ and α is a value between 0 and 1. The latter value is used to balance the importance of each objective function.

4 Experimental results

This section presents the computational results for the ATFM-TCA formulations of Section 3. The tests have been performed in a laboratory environment by using real-world instances of FCO TCA. The ATFM-TCA solutions are computed by using CPLEX MIP solver 12.0. The experiments are executed on processor Intel Core 2 Duo E6550 (2.33 GHz), 2 GB of RAM and Windows XP operative system.

Table 1 gives information on the 20 ATFM-TCA instances considered. We generated randomly 20 disturbed scenarios, with delays up to 5 minutes.

Column 1 reports the time period of traffic prediction (in minutes), Columns 2-3 the number of landing and departing aircraft, Columns 4-5 the maximum entrance and unavoidable delays (in seconds), and Column 6 the total number of aircraft routes considered.

Table 1: Fiumicino airport instances

Time Period (min)	Landing Aircraft	Departing Aircraft	Max Entrance Delay (sec)	Max Unavoid. Delay (sec)	Aircraft Routes
0–30	16	4	294	160	36

We next show the results obtained for single and combined objective functions.

4.1 Single Objective Functions

Table 2 reports the results for MT (Columns 2-5) and TTTS (Columns 6-9). Row 1 gives the objective function (*OBJ*) used to solve the 20 ATFM-TCA instances. Row 2 presents the studied ATFM-TCA problems: *Scheduling* is the problem with aircraft routes fixed off-line (fixed so that the workload of the runways is well balanced and there is no conflict at runways when aircraft are on time), and *Routing* is the problem with routing flexibility. Rows 3–5 are the results obtained with a maximum computation time of 1 minute: Row 3 is the number of optimal solutions (*Optimality*) found by CPLEX within a given time limit, Row 4 is the objective function value (*UB*, in seconds), Row 5 is the Lower Bound on the optimal solution given by CPLEX (*LB*, in seconds). Rows 6–8 present the same type of information of Rows 3–5 but a time limit of 60 minutes is now given to CPLEX. Rows 9–10 give the values obtained for

the other performance indicators for the solutions obtained after 1 minute of computation. For each problem (Scheduling or Routing) we report the average results obtained by CPLEX plus the gap (in %) between the value obtained for each performance indicator and the corresponding best known value for the corresponding problem.

Table 2: ATFM-TCA solutions computed for single objective functions

OBJ	**MT**				**TTTS**			
Problem	Scheduling	Gap	Routing	Gap	Scheduling	Gap	Routing	Gap
Time Limit of 1 Minute								
Optimality	20		2		1		0	
UB (sec)	61		89		13189		13026	
LB (sec)	61		0		12338		11079	
Time Limit of 60 Minutes								
Optimality	20		20		20		0	
UB (sec)	61		24		13183		12521	
LB (sec)	61		24		13183		11661	
MT (sec)	-	-	-	-	315	421.1	184	179.1
TTTS (sec)	13414	1.8	13790	9.7	-	-	-	-

When evaluating the number of optimal solutions found in Table 2, it is clear that MT minimization is easier to solve than TTTS minimization. In fact, MT is influenced by fewer aircraft compared to TTTS.

From a comparison of the results obtained in 60 seconds with the ones computed in 60 minutes, the solution found for the scheduling problem in the initial 60 seconds it is (near)optimal. The routing problem is by far more difficult to solve due to the larger number of variables. However, the optimal solution for MT is always found in one hour computation. Differently, a small gap still exists for the routing problem with TTTS minimization.

We now look at the gaps between the best solutions computed after 1 minute for one objective function and evaluated with the other. The solutions found for the minimization of TTTS have poor performance in terms of MT, while the solutions found for MT have small gaps ($< 6\%$) regarding TTTS. This trend can be justified by a low correlation between the travel time and the aircraft delay minimization, since several aircraft may satisfy their scheduled time even in presence of delays. This is especially true for landing aircraft that have, by construction, a large recovery time in their assigned time window of departure.

4.2 Combined Objective Functions

To study the combined approach, we analyze the average results obtained for the 20 ATFM-TCA instances by varying the parameter α in the window $[0; 0.1; ...; 0.9; 1]$, as reported in

Figure 2. Also, we use as δ and γ the UB values computed by CPLEX with one-hour computation time, see Table 2. In this section, we give a computation time of 1 minute (2 minutes) to the scheduling (routing) problem.

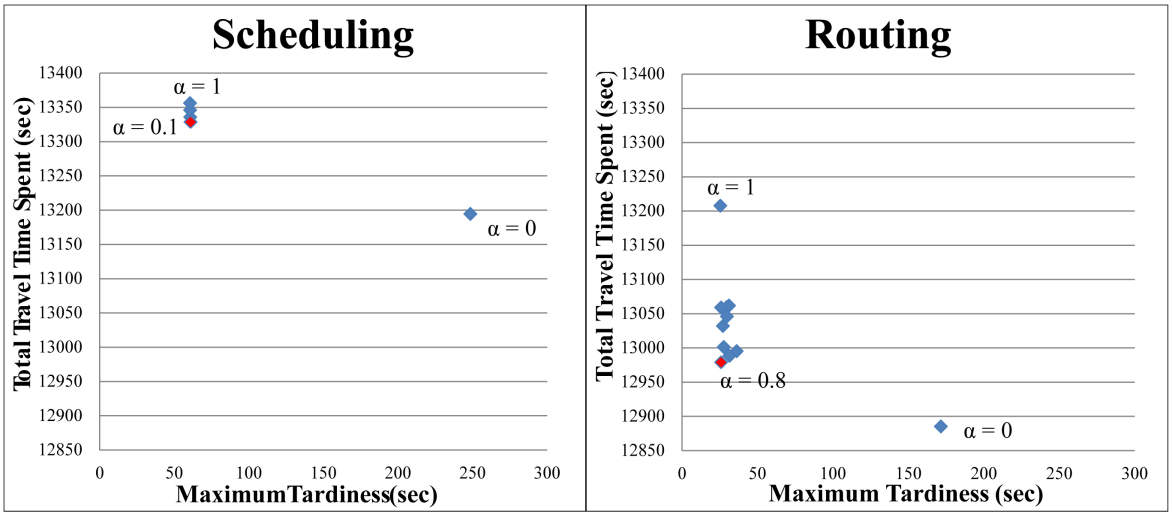


Figure 2: MT-TTTS for various values of α

A good trade-off between the single objective functions is obtained for $\alpha = 0.1$ in the scheduling case and $\alpha = 0.8$ in the routing one. Specifically, MT is stable around its best value, in particular for fixed routing. TTTS deteriorates quickly when it loses weight in the objective function.

We conclude that no one objective function outperforms the others in terms of both performance indicators. In fact, the solutions resulting from the combination of the single objective functions have the drawback to deteriorate the performance of at least an indicator.

5 Conclusions and further research

This paper presents microscopic formulations of the ATFM-TCA problem with two objective functions of practical interest: minimization of the largest delay and the total travel time spent in the TCA. Experimental results show the existence of relevant gaps between the different ATFM-TCA solutions computed by CPLEX. From our experiments, the combination of the two objective functions offers a good trade-off between the performance indicators. In general, we believe that this work moves the interest of researchers and practitioners in paying more attention to the various performance indicators and thus on the inherent multi-objective nature of the ATFM-TCA problem.

Ongoing research is dedicated to the study of additional objective functions and more severe traffic disturbances, including temporary blocked runways. Future research will also be focused to the development of real-time scheduling and routing algorithms for multi-objective optimization models in order to reduce the optimality gap found by CPLEX.

References

[And13] G. ANDREATTA, L. CAPANNA, L. DE GIOVANNI, M. MONACI, and L. RIGHI: "Efficiency and Robustness in Integrated Airport Apron, a Support Platform for Intelligent Airport Ground Handling". In: *Journal of Intelligent Transportation Systems* IN PRESS.3 (2013), pp. 1–67.

[Bar12] C. BARNHART, D. FEARING, A. ODONI, and V. VAZE: "Demand and capacity management in air transportation". In: *EURO Journal of Transportation and Logistics.* 1.1-2 (2012), pp. 135–155.

[Ben11] J. BENNELL, M. MESGARPOUR, and C. POTTS: "Airport runway scheduling". In: *4OR - Quarterly Journal of Operations Research* 9.2 (2011), pp. 115–138.

[Ber11] D. BERTSIMAS, G. LULLI, and A. ODONI: "An integer optimization approach to large-scale air traffic flow management". In: *Operations Research* 59.1 (2011), pp. 211–227.

[Bia06] L. BIANCO, P. DELL'OLMO, and S. GIORDANI: "Scheduling models for air traffic control in terminal areas". In: *Journal of Scheduling* 9.3 (2006), pp. 180–197.

[Cas11] L. CASTELLI, R. PESENTI, and A. RANIERI: "The design of a market mechanism to allocate Air Traffic Flow Management slots". In: *Transportation Research Part C* 19.5 (2011), pp. 931–943.

[Chu10] A. CHURCHILL, D. LOVELL, and M. BALL: "Flight delay propagation impact on strategic air traffic flow management". In: *Transportation Research Records* 2177 (2010), pp. 105–113.

[Cla10] J. CLAUSEN, A. LARSEN, and J. LARSEN: "Disruption management in the airline industry–Concepts, models and methods". In: *Computers and Operations Research* 35.5 (2010), pp. 809–821.

[DAr07] A. D'ARIANO, D. PACCIARELLI, and M. PRANZO: "A branch and bound algorithm for scheduling trains in a railway network". In: *European Journal of Operational Research* 183.2 (2007), pp. 643–657.

[DAr10] A. D'ARIANO, P. D'URGOLO, D. PACCIARELLI, and M. PRANZO: "Optimal Sequencing of aircrafts Take-Off and Landing at a Busy Airport". In: *Proceedings of the 13th International IEEE Conference on Intelligent Transportation Systems*. Vol. 33. Series Transportation Research. Madeira Island, Portugal, 2010, pp. 1569–1574.

[DAr12] A. D'ARIANO, M. PISTELLI, and D. PACCIARELLI: "Aircraft retiming and rerouting in vicinity of airports". In: *IET Intelligent Transport Systems* 6.4 (2012), pp. 433–443.

[Kim11] J. KIM, A. KRÖLLER, J. MITCHELL, and G. SABHNANI: "Scheduling Aircraft to Reduce Controller Workload". In: *Proceedings of the 9th Workshop on Algorithmic Approaches for Transportation Modeling. Optimization. and Systems*. Vol. 33. Series Transportation Research. ATMOS. Copenhagen, Denmark, 2011, pp. 1–4.

[Koh07] N. KOHL, A. LARSEN, J. LARSEN, A. ROSS, and S. TIOURINE: "Airline disruption management–Perspectives, experiences and outlook". In: *Journal of Air Transport Management* 13.3 (2007), pp. 149–162.

[Kuc00] J. KUCHAR and L. YANG: "A Review of Conflict Detection and Resolution Modeling Methods". In: *IEEE Trans. on Intelligent Transportation Systems* 4.1 (2000), pp. 179–189.

[Mas02] A. MASCIS and D. PACCIARELLI: "Job shop scheduling with blocking and no-wait constraints". In: *European Journal of Operational Research* 143.3 (2002), pp. 498–517.

[Pel12] P. PELLEGRINI, L. CASTELLI, and R. PESENTI: "Metaheuristic algorithms for the simultaneous slot allocation problem". In: *IET Intelligent Transport System* 6.4 (2012), pp. 453–462.

[Sam13a] M. SAMÀ, A. D'ARIANO, and D. PACCIARELLI: "Rolling Horizon Approach for Aircraft Scheduling in the Terminal Control Area of Busy Airports". In: *Procedia - Social and Behavioral Sciences* 80 (2013). 20th International Symposium on Transportation and Traffic Theory (ISTTT 2013), pp. 531–552. ISSN: 1877-0428. DOI: http://dx.doi.org/10.1016/j.sbspro.2013.05.029. URL: http://www.sciencedirect.com/science/article/pii/S1877042813009981.

[Sam13b] M. SAMÀ, P. D'ARIANO, A. D'ARIANO, and D. PACCIARELLI: "Scheduling models for optimal aircraft traffic control at busy airports: tardiness, priorities, equity and violations considerations". In: *Tech. Rep. RT-DIA-205-2013, Dipartimento di Informatica e Automazione, "RomaTre"* 4.1 (2013), pp. 1–24.

Corresponding author: Andrea D'Ariano, Università degli Studi Roma Tre, Dipartimento di Ingegneria, via della Vasca Navale, 79 - 00146 Rome, Italy, phone: +39 06 5733 3456, e-mail: a.dariano@dia.uniroma3.it

A Mixed-Integer Optimal Control Approach for Aircraft Landing Model

Konstantin D. Palagachev[1], Matthias Rieck[2], Matthias Gerdts[1]
[1]Universität der Bundeswehr München
[2]Technische Universität München

Abstract

This paper presents a mixed-integer optimal control method applied to aircraft landing problem. The optimization task is to minimize fuel consumption under speed constraints depending on the discrete valued control, representing the flaps position of the aircraft.

Keywords: Mixed-Integer Optimal Control, Vanishing Constraints, Aircraft Landing Model

1 Introduction

The rapid growth of worldwide air travels and the constant demand of efficiency require fuel consumption reduction. Modern transport aircrafts are designed to operate in cruise flights. Nevertheless, efficiency of the aircraft in configurations away from the optimal one, such as taking off and landing can be further optimized.

In this paper we propose an aircraft model aiming to reduce fuel consumption during landing. Due to speed reduction during the manoeuvre, flaps have to be extended in order to increase the lift force acting on the aircraft. A continuous approximation of the flaps position was already studied in [Lau11], while in [Lau10] automatic flaps for decelerations were investigated. Our approach is different from that proposed in [Lau11] and [Lau10], since we consider flaps as control, which is forced to assume only a discrete number of values. This leads to a mixed-integer optimal control problem (optimal control problem in which both discrete and continuous valued controls appear). The obtained optimal trajectory can be used as a reference solution for adaptive controllers as the ones proposed in [H C10] and [Dal13]

2 Problem Formulation

An abstract formulation of our mixed-integer optimal control problem is given as follows:

(MIOCP) Minimize $\varphi(x(t_f))$, with respect to $x \in W^{1,\infty}([t_0,t_f],\mathbb{R}^{n_x})$, $u \in L^\infty([t_0,t_f],\mathbb{R}^{n_u})$ and $v \in L^\infty([t_0,t_f],\mathbb{R})$, subject to

$$\begin{aligned} \dot{x}(t) - f(x(t),u(t),v(t)) &= 0 \qquad \text{a.e. } t \in [t_0,t_f], \\ g(x(t),u(t),v(t)) &\le 0 \qquad \text{a.e. } t \in [t_0,t_f], \\ \psi(x(t_0),x(t_f)) &= 0, \\ v(t) &\in \{v_1,\dots,v_M\} \qquad \text{a.e. } t \in [t_0,t_f]. \end{aligned} \tag{1}$$

For $n \in \mathbb{N}$, the space $L^\infty([t_0,t_f],\mathbb{R}^n)$ consists of all measurable functions $f : [t_0,t_f] \to \mathbb{R}^n$ with

$$\|f\|_\infty = \operatorname*{ess\,sup}_{t\in[t_0,t_f]} \|f(t)\| < \infty$$

where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^n$. The space $W^{1,\infty}([t_0,t_f],\mathbb{R}^n)$ consists of all absolutely continuous functions $f : [t_0,t_f] \to \mathbb{R}^n$ with

$$\|f\|_{1,\infty} = \max\{\|f\|_\infty, \|\dot{f}\|_\infty\} < \infty$$

where $\dot{f}$ is the derivative of f. We assume that the functions $\varphi : \mathbb{R}^{n_x} \to \mathbb{R}$, $f : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \times \mathbb{R} \to \mathbb{R}^{n_x}$, $g : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \times \mathbb{R} \to \mathbb{R}^{n_g}$ and $\psi : \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_\psi}$ are continuously differentiable with respect to all variables.

There are several ways to solve (MIOCP) numerically. One way is to formulate and solve necessary optimality conditions given by the Pontryagin maximum principle, as done in [Iof79, Theorem 1, p. 234]. However, this method is problem depending and turns out to be complex for complicated problems. An alternative approach proposed in [Ger05] is to discretize the problem and solve a large scale finite dimensional mixed-integer mathematical program using Branch&Bound method, however this method becomes inefficient as the number of discretization points grows. In this paper, we follow a different approach based on a suitable time transformation. The first formulation of this method appears in literature in [Dub65, Section 7, p. 47], while in [Ger06] its application to mixed-integer optimal control problems is considered.

Variable Time Transformation

Let us define for every $n \in \mathbb{N}$ the following grid, which we will call *major grid* $\mathbb{G}_N := \left\{ t_i = t_0 + ih \mid i = 0,\dots,N,\ h = \frac{t_f - t_0}{N} \right\}$. We need now to partition each time interval of the major grid $[t_i, t_{i+1})$ into M disjoint subintervals $[\tau_{i,j}, \tau_{i,j+1})$, where M is the number of values assumed by the discrete control v. Thus, we obtain the *minor grid* defined by $\mathbb{G}_{N,M} := \left\{ \tau_{i,j} = t_i + j\frac{h}{M} \mid i = 0,\dots,N-1,\ j = 0,\dots,M \right\}$ under the convention $\tau_{i,M} = \tau_{i+1,0} = t_{i+1}$ for every $i = 0,\dots,N-1$. Let us now define the piecewise constant function on the minor grid $v_{\mathbb{G}}(\tau) := v_j$ for every $\tau \in [\tau_{i,j-1}, \tau_{i,j})$, $i = 0,\dots,N-1$, $j = 1,\dots,M$, and the time

transformation

$$t(\tau) := t_0 + \int_{t_0}^{\tau} w(s)ds \qquad \forall\, \tau \in [t_0, t_f] \tag{2}$$

where $w \in L^\infty([t_0, t_f], \mathbb{R})$. The following constraints have to be imposed on the function w :

$$w(s) \geq 0 \qquad \text{a.e. } s \in [t_0, t_f] \quad \text{and} \quad \int_{t_i}^{t_{i+1}} w(s)ds = h \;\forall\, i = 1, \ldots, N. \tag{3}$$

Note that (3) prevent the time of running backward and keep the length of major grid time intervals fixed. In this way we control the length of the minor time intervals $[t(\tau_{i,j-1}), t(\tau_{i,j}))$, while keeping fixed the length of the major time intervals $[t(t_i), t(t_{i+1}))$. Under constraints (3) , the transformation $t(\cdot)$ maps $[t_0, t_f]$ onto itself. Furthermore, it is absolutely continuous and its derivative is given by $\frac{dt}{d\tau}(\tau) = w(\tau)$ for almost every $\tau \in [t_0, t_f]$. Note that $t(\cdot)$ is not invertible, in fact all time intervals in which $w \equiv 0$ are mapped into a single point. Anyway, it becomes invertible under the convention

$$t^{-1}(s) = \inf\{\tau \mid t(\tau) = s\}. \tag{4}$$

Note that every feasible discrete control $v(\cdot)$ of (MIOCP) can be written as

$$v(s) = v_{\mathbb{G}}(t^{-1}(s)) \tag{5}$$

Using now (2) in (MIOCP), we obtain the following transformed mixed-integer optimal control problem

(TMIOCP) Minimize $\varphi(x(t_f))$, with respect to $x \in W^{1,\infty}([t_0, t_f], \mathbb{R}^{n_x})$, $u \in L^\infty([t_0, t_f], \mathbb{R}^{n_u})$ and $w \in L^\infty([t_0, t_f], \mathbb{R})$, subject to

$$\dot{x}(\tau) - w(\tau) f(x(\tau), u(\tau), v_{\mathbb{G}}(\tau)) = 0 \qquad \text{a.e. } \tau \in [t_0, t_f], \tag{6}$$

$$w(\tau) g(x(\tau), u(\tau), v_{\mathbb{G}}(\tau)) \leq 0 \qquad \text{a.e. } \tau \in [t_0, t_f], \tag{7}$$

$$w(\tau) \geq 0 \qquad \text{a.e. } \tau \in [t_0, t_f], \tag{8}$$

$$\int_{t_i}^{t_{i+1}} w(\tau)d\tau = h \qquad \forall\, i = 0, \ldots, N-1, \tag{9}$$

$$\psi(x(t_0), x(t_f)) = 0. \tag{10}$$

Note that we have obtained (7) multiplying (1) by the function w. This makes sense, since $g(x(\tau), u(\tau), v_{\mathbb{G}}(\tau)) \leq 0$ has only to be considered on those intervals on which $w(\tau) > 0$.

Solving (TMIOCP) by direct discretization method on the minor grid $\mathbb{G}_{N,M}$ leads to the optimal solutions $x^*,\ u^*,\ w^*$. Approximated solutions of (MIOCP) are given by

$$x(s) := x^*(t^{-1}(s)), \quad u(s) := u^*(t^{-1}(s)), \quad v(s) := v_{\mathbb{G}}(t^{-1}(s))$$

where $t(\tau) = t_0 + \int_{t_0}^{\tau} w^*(s)ds$, while $t^{-1}(s)$ is given by (4).

2.1 Numerical Solution

In order to solve (TMIOCP) numerically, we consider continuous and piecewise linear approximation of the functions x, u and w on the minor grid: $x(\tau) := x_{i,j},\ x_{i,M} = x_{i+1,0},\ x(t_f) = x_{N,0},\ u(\tau) := u_{i,j},\ w(\tau) = w_{i,j}$ for every $\tau \in [\tau_{i,j-1}, \tau_{i,j})$, where $i = 0, \ldots, N-1,\ j = 1, \ldots, M$. The system of differential equations (6) is discretized by an $s-$staged Runge-Kutta method with coefficients b_k, c_k, $a_{k\mu}$, $1 \le k, \mu \le s$, on the minor grid $\mathbb{G}_{N,M}$:

$$x_{i,j+1} - x_{i,j} - \frac{h}{M} \sum_{k=1}^{s} b_k \xi_k^{i,j} = 0 \quad i = 0, \ldots, N-1,\ j = 0, \ldots, M-1$$

where $\xi_k^{i,j} = w_{i,j} f(x_{i,j} + \frac{h}{M} \sum_{\mu=1}^{s} a_{k\mu} \xi_\mu^{i,j}, u_{i,j}, v_j)$. Furthermore, (7)-(10) become

$$\begin{aligned} &w_{i,j} g(x_{i,j}, u_{i,j}, v_j) \le 0, \quad && w_{i,j} \ge 0, \\ &\textstyle\sum_{j=1}^{M} w_{i,j} = M, && \psi(x_{0,0}, x_{N,0}) = 0. \end{aligned}$$

In this way, after discretization on the minor grid, (TMIOCP) is transformed into the following mathematical program

(MPVC) Minimize $\varphi(x_{N,0})$, with respect to $x_{i,j}, x_{N,0} \in \mathbb{R}^{n_x}$, $u_{i,j}\mathbb{R}^{n_u}$ and $w_{i,j} \in \mathbb{R}$, subject to

$$\begin{aligned} x_{i,j+1} - x_{i,j} - \frac{h}{M} \sum_{k=1}^{s} b_k \xi_k^{i,j} &= 0, \\ w_{i,j} g(x_{i,j}, u_{i,j}, v_j) &\le 0, && (11) \\ w_{i,j} &\ge 0, && (12) \\ \sum_{j=1}^{M} w_{i,j} - M &= 0, \\ \psi(x_{0,0}, x_{N,0}) &= 0 \end{aligned}$$

for all $i = 0, \ldots, N-1$ and $j = 1, \ldots, M$.

We would like to emphasize the particular structure of (MPVC), more precisely constraints (11) and (12). Such a type of constraints are called vanishing, since (11) is automatically satisfied for every $i = 0, \ldots, N-1$ and $j = 1, \ldots, M$ for which (12) is active (i.e. $w_{i,j} = 0$). A first formal treatment has been done in [Ach08], where also necessary optimality conditions had been obtained. An extensive overview of the argument can be found in [Hoh09].

In general, it is not possible to solve (MPVC) directly using sequential quadratic programming (SQP) method. This is due to the combinatorial nature of constraints (11)-(12). This is the reason why (MPVC) have to be approximated by a non-linear mathematical program, whose solution converges to the solution of the original one. So far, two main techniques are used in literature, a smoothing and relaxation approach. Formulation and convergence properties for both of them can be found in [Hoh09, Chapter 10]. We will follow the second one. Let us relax constraint (11) by the positive parameter $\tau > 0$, i.e $w_{i,j} g(x_{i,j}, u_{i,j}, v_j) \le \tau$.

Substituting the last inequality in (MPVC), we obtain the relaxed non-linear mathematical program

(NLP(τ)) Minimize $\varphi(x_{N,0})$ with respect to $x_{i,j}, x_{N,0} \in \mathbb{R}^{n_x}$, $u_{i,j}\mathbb{R}^{n_u}$ and $w_{i,j} \in \mathbb{R}$, subject to

$$\begin{aligned} x_{i,j+1} - x_{i,j} - \tfrac{h}{M}\textstyle\sum_{k=1}^{s} b_k \xi_k^{i,j} = 0, \quad & \textstyle\sum_{j=1}^{M} w_{i,j} - M = 0, \\ w_{i,j} \geq 0, \quad & w_{i,j} g(x_{i,j}, u_{i,j}, v_j) \leq \tau, \\ \psi(x_{0,0}, x_{N,0}) = 0 & \end{aligned}$$

for all $i = 0, \ldots, N-1$ and $j = 1, \ldots, M$.

3 Aircraft Model

In order to optimize a three dimensional approach trajectory a suitable 3DOF model of the aircraft dynamics is necessary. The model consists of position (in a local NED-North East Down frame), translatory (kinematic frame) as well as a fuel flow differential equations. Wind influence is not taken into account. Therefore the kinematic states are equal to the aerodynamic states. A list of all the states can be seen below.

Parameter	**Symbol**	**Unit**
x-Position in local NED Frame	x_O	$[m]$
y-Position in local NED Frame	y_O	$[m]$
z-Position in local NED Frame	z_O	$[m]$
Speed	V	$[m/s]$
Course Angle	χ	$[rad]$
Climbing Angle	γ	$[rad]$
Mass	m	$[kg]$
Bank Angle	μ	$[rad]$

In the same way we list the controls applied on the aircraft:

Parameter	**Symbol**	**Unit**	**Range**
Lift Coefficient	C_L	$[-]$	$[0, 2.68]$
Side Force Coefficient	C_Q	$[-]$	$[0, 0]$
Bank Angle Time Derivative	$\dot{\mu}$	$[rad/s]$	$[-0.2618, +0.2618]$
Thrust Lever Position	δ_T	$[-]$	$[0, 1]$

Please note that for the bank angle of the aircraft the time derivative is commanded, which otherwise would introduce strong oscillations in the OCP solution. Therefore the bank control μ is an additional state in the aircraft dynamics. Since we want to prevent any angle of sideslip we set the lower and upper bound for the side force coefficient to zero.

Calculating the time derivative of the position is straight forward. Position is given relative to an origin in a local NED frame (please note that the z-axis points downwards):

$$\dot{x} = V \cdot \cos\chi \cdot \cos\gamma, \qquad \dot{y} = V \cdot \sin\chi \cdot \cos\gamma, \qquad \dot{z} = -V \sin\gamma.$$

Formulas (13) describe the translatory differential equations of the aircraft:

$$\dot{V} = \frac{X^T}{m} - \sin\gamma \cdot g, \qquad \dot{\chi} = \frac{Y^T}{m \cdot V \cdot \cos\gamma}, \qquad \dot{\gamma} = -\frac{Z^T}{m \cdot V} - \frac{\cos\gamma \cdot g}{V}. \tag{13}$$

The forces acting on an aircraft are aerodynamic (X^A, Y^A, Z^A), propulsive (P) and ground forces. Since we will not simulate an actual touch down and our final boundary condition is the final fix point for the runway ground forces are not taken into account. Furthermore the propulsive force is regarded to act along the x-axis of the kinematic frame.

$$X^T = X^A + P, \qquad Y^T = Y^A, \qquad Z^T = Z_A.$$

The aerodynamic forces acting on the aircraft are the lift L, the side force Q and the drag D. The drag is calculated using the symmetric quadratic polar. All coefficients used in the model are taken from the Base of Aircraft Dataset (BADA) from Eurocontrol [BAD39].

$$\begin{aligned} q &= \tfrac{\rho}{2} \cdot V^2, & L &= q \cdot S \cdot C_L, \\ Q &= q \cdot S \cdot C_Q, & D &= q \cdot S \cdot (C_{D0} + C_{D2} \cdot C_L^2). \end{aligned}$$

The forces calculated above are the aerodynamic forces if the x-axis of the aerodynamic frame points to the back of the aircraft. Since the aerodynamic frame here points to the aircraft nose, a correction needs to be applied. The aerodynamic forces $(X^A)_A, (Y^A)_A, (Z^A)_A$ are:

$$(X^A)_A = -D, \qquad (Y^A)_A = Q, \qquad (Z^A)_A = -L.$$

However, the aerodynamic forces are needed in the kinematic frame K (X^A, Y^A, Z^A,). To achieve this a rotation has to be applied around the x-axis

$$X^A = (X^A)_A, \quad Y^A = (Y^A)_A \cdot \cos\mu - (Z^A)_A \cdot \sin\mu, \quad Z^A = (Y^A)_A \cdot \sin\mu + (Z^A)_A \cdot \cos\mu.$$

In order to model the propulsive force the Aircraft Noise and Performance database (ANP) is used. The available thrust is dependent on the aircraft speed V, the altitude h (which influences the air density ρ and pressure p) and the thrust lever position δ_T. First the Calibrated Air Speed (CAS) is calculated by $V_{CAS} = \sqrt{\frac{\rho}{\rho_0}} \cdot V$. Then the maximum corrected netto thrust of one engine is calculated which is multiplied with the thrust lever position to obtain the corrected netto thrust of one engine. Please note that the symbols $[x2y]$ refer to the unit conversion factor of unit x to unit y.

$$\begin{aligned} P^{corr,max,eng,lbf} &= E + F \cdot V_{CAS} \cdot [ms2kt] \\ &\quad + G_a \cdot h \cdot [m2ft] + G_b \cdot h^2 \cdot [m2ft]^2 \\ P^{corr,max,eng} &= P^{corr,max,eng,lbf} \cdot [lbf2N] \\ P^{corr,eng} &= P^{corr,max,eng} \cdot \delta_T \end{aligned}$$

Afterwards the current thrust can be calculated from the corrected thrust using the air pressure correction. Finally the current thrust is multiplied with the number of engines n_{eng} to obtain the overall thrust:

$$P^{eng} = P^{corr,max,eng} \cdot \frac{p}{p_0}, \qquad P = n_{eng} \cdot P^{eng}.$$

To take into account the mass change of the aircraft throughout its flight, the fuel flow to the engines and therefore the change of mass has to be modelled. The fuel flow model is taken from the BADA dataset. This results in the following differential equation for the aircraft mass: $\dot{m} = -f_{flow}$. The fuel flow is mainly influenced by the thrust lever position δ_T. However if $\delta_T = 0$, the engine still has an idle fuel flow f_{min}. Therefore the fuel flow is interpolated between the minimum and maximum fuel flow

$$f_{flow} = f_{min} + \delta_T \cdot (f_{max} - f_{min}),$$

where f_{min} is dependent on the fuel flow coefficients C_{f3} and C_{f4} as well as the geopotential altitude H_G, while R_e represents the earth radius.

$$f_{min} = C_{f3} \cdot (1 - \frac{H_G}{C_{f4}}), \qquad H_G = \frac{R_e \cdot h}{R_e + h}, \qquad h = -z.$$

In order to create a realistic aircraft trajectory which is comfortable for the pilots and passengers a few additional path constraints have to be introduced to the problem. First of all the bank angle is limited to $\pm 32°$. Secondly the aircraft must not perform high g manoeuvres. Therefore the load factor in z-direction is limited to $n_z \in [0.8, 1.2]$. Finally for the approach to the runway both, passengers as well as pilot, prefer trajectories in which the aircraft altitude and the speed is reduced only. This results in the following path constraints: $\dot{V} \leq 0, \quad \dot{z} \geq 0$.

On any aircraft the speed flight envelope (the speed range the aircraft is allowed to fly) is dependent on the flap position. Since the flap position depends on the control, the speed range has to be adapted dynamically. We use Vanishing Constraint as mentioned above to account for them. For every discretization step and every discrete control value two Vanishing Constraints have to be defined (one for each lower an upper bound):

$$VC_{V_{min}} = w_i \cdot (V_{min} - V), \qquad VC_{V_{max}} = w_i \cdot (V - V_{max}).$$

4 Switching Costs

With the Variable Time Transformation the optimizer is able to alter the discrete controls at every time step. However the optimal solution may have a discrete control switch at every time step. Since in our case flaps are modelled by the discrete control, they should be changed as few times as possible. Therefore to limit the switches, a speed dependent cost function is introduced in order to penalize discrete control switches. The idea for the speed dependent

switching cost lies in the fact that the flaps need to be changed whenever the speed reaches a flap dependent limit. If this is the case the switching cost becomes zero which enables the optimizer to try different discrete controls. The function which models the speed dependent cost is defined by multiple hyperbolic tangent functions added together (see Fig.1). A switch

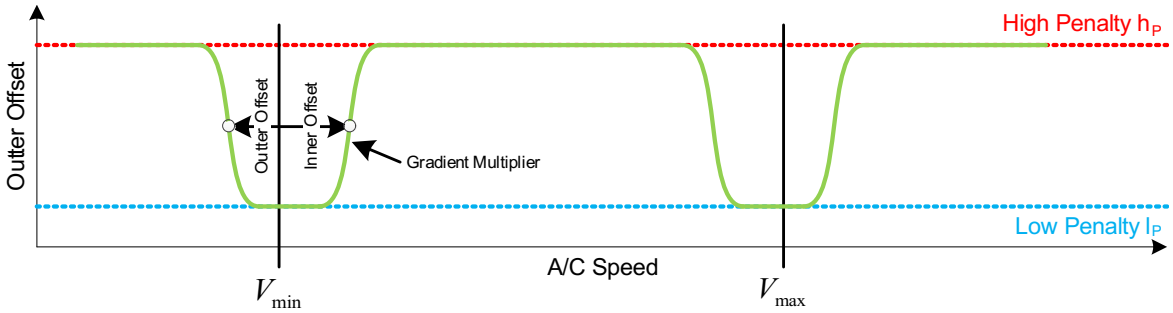


Figure 1: Speed Dependent Switching Cost

is detected by multiplying the corresponding w values from the Variable Time Transformation (see (14) for speed dependent penalty for switch from cruise to approach)

$$J_P = p(V) \cdot w_{i,CR} \cdot w_{i+1,APR}. \tag{14}$$

5 Optimization

In the example optimization the approach trajectory from Munich airport on runway 08L from MIKE VOR is optimized. In our case the approach trajectory is given by four waypoints. For the second and third waypoint the x, y, z-coordinates have been relaxed since the pilot is allowed to deviate from these points. The discrete controls were initialized equally, which

Location	χ	γ	V
$48°34''N, 11°36''E, 5000ft$	$227°$	$0°$	$108\frac{m}{s}$
$48°25''N, 11°22''E, 5000ft$	–	–	–
$48°20''N, 11°27''E, 5000ft$	–	–	$80\frac{m}{s}$
$48°21''N, 11°41''E, 2320ft$	$82°$	$-3°$	$[58, 65]\frac{m}{s}$

Table 1: Boundary Condition for Optimal Control Problem

means a switch at every discretization step. This way the optimizer is in charge of choosing the correct switching sequence itself. As can be seen in Fig.2 the resulting trajectory follows all waypoint with straight flights between them. The aircraft speed over time plot is shown in Fig.3. It can be seen that the aircraft switches from cruise to approach at around 500 seconds. The switch to the landing configuration occurs approximately 14 seconds later. The overall time lies at 880.2 seconds.

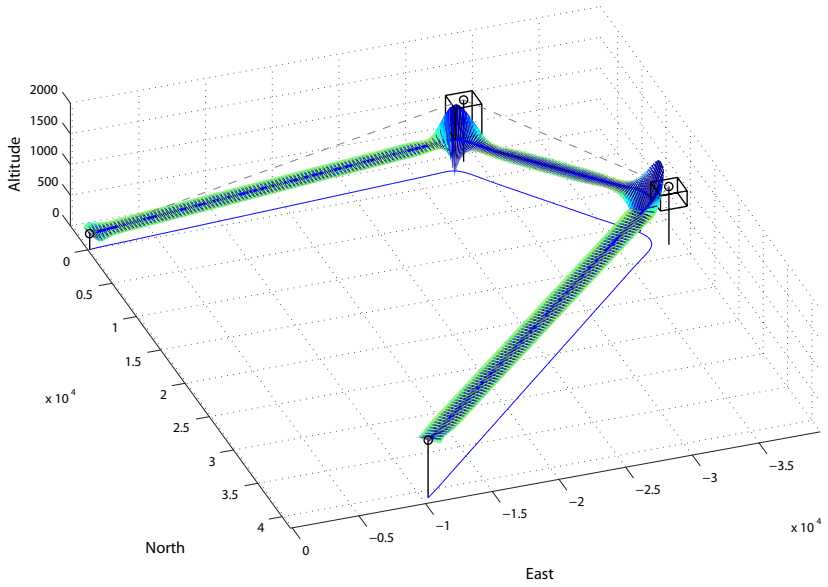


Figure 2: Time Optimal Trajectory

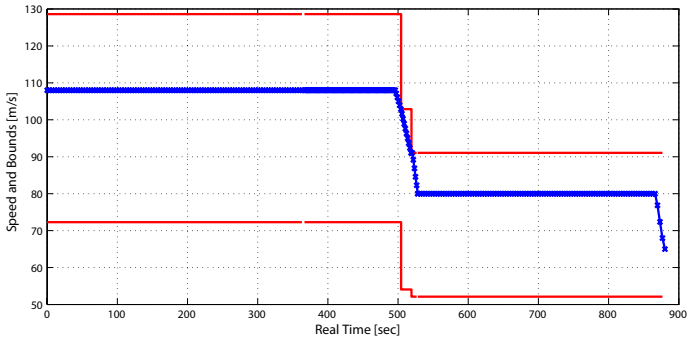


Figure 3: Speed

6 Conclusions

In summery, we propose a method for solving mixed-integer optimal control problems, capable to deal with constraints depending on the discrete valued control. Since the combinatorial nature of the problem vanishes after reformulation, gradient based methods have been used in order to obtain optimal solution. We apply the method on aircraft landing model, in which fuel reduction during the manoeuvre is aimed. Flaps position is considered as additional discrete control and speed constraints depending on flaps are considered.

References

[Ach08] W. ACHTZIGER and C. KANZOW: "Mathematical Programs with Vanishing Constraints: Optimality Conditions and Constraint Qualifications". In: *Mathematical Programming* 114.1 (July 2008), pp. 69–99. ISSN: 0025-5610.

[BAD39] *User Manual for the Base of Aircraft Data (BADA)*. 3.9. Eurocontrol.

[Dal13] L. DALLDORF, R. LUCKNER, and R. REICHEL: "A Full-Authority Automatic Flight Control System for the Civil Airborne Utility Platform S15 – LAPAZ". In: *CEAS 2nd EuroGNC 2013, Conference in Guidance, Navigation and Control in Aerospace, Delft*. Apr. 2013.

[Dub65] A. DUBOVITSKII and A. MILJUTIN: "Extremum Problems in the Presence of Restrictions". In: *U.S.S.R. Computational Mathematics and Mathematical Physics* 5.3 (1965), pp. 1–80.

[Ger05] M. GERDTS: "Solving Mixed-Integer Optimal Control Problems by Branch and Bound: a case study from automobile testdriving with gear shift". In: *Optimal Control Applications and Methods* 26.1 (2005), pp. 1–18.

[Ger06] M. GERDTS: "EA Variable Time Transformation Method for Mixed-Integer Optimal Control Problems". In: *Optimal Control Applications and Methods* 27.3 (2006), pp. 169–182.

[H C10] Y. C. H. CHAO and Y. CHEN: "Autopilots for Small Unmanned Aerial Vehicles: A Survey". In: *International Journal of Control, Automation, and Systems* 8.1 (2010), pp. 36–44.

[Hoh09] T. HOHEISEL: "Mathematical Programs with Vanishing Constraints". PhD thesis. Wuerzburg, Germany: University of Wuerzburg, 2009. URL: `http://www.mathematik.uni-wuerzburg.de/~hoheisel/diss.pdf`.

[Iof79] A. IOFFE and V. TIHOMIROV: *Theory of Extremal Problems (Studies in mathematics and its applications)*. Elsevier Science Ltd, 1979.

[Lau10] M. LAUTERBACH and R. LUCKNER: "Kontrollierte Verzoegerung im Landeanflug mittels automatischer Hochauftriebshilfen". In: *Deutscher Luft- und Raumfahrtkongress, Hamburg*. Aug. 2010.

[Lau11] M. LAUTERBACH and R. LUCKNER: "Vergleich verschiedener Regelstrategien fuer die stufenlose Verstellung der Hinterkantenklappen im Landeanflug". In: *Deutscher Luft- und Raumfahrtkongress, Bremen*. Sept. 2011.

Corresponding author: Konstantin D. Palagachev, Universität der Bundeswehr München, Institute of Mathematics and Applied Computing, 85577 Neubiberg/München, Germany, phone: +49 89 6004 3082, e-mail: konstantin.palagachev@unibw.de

Potential Fields in Maritime Anomaly Detection

Ewa Osekowska, Stefan Axelsson, Bengt Carlsson
Blekinge Institute of Technology

Abstract

This paper presents a novel approach for pattern extraction and anomaly detection in maritime vessel traffic, based on the theory of potential fields. Potential fields are used to represent and model normal, i.e. correct, behaviour in maritime transportation, observed in historical vessel tracks. The recorded paths of each maritime vessel generate potentials based on metrics such as geographical location, course, velocity, and type of vessel, resulting in a potential-based model of maritime traffic patterns.

A prototype system STRAND, developed for this study, computes and displays distinctive traffic patterns as potential fields on a geographic representation of the sea. The system builds a model of normal behaviour, by collating and smoothing historical vessel tracks. The resulting visual presentation exposes distinct patterns of normal behaviour inherent in the recorded maritime traffic data. Based on the created model of normality, the system can then perform anomaly detection on current real-world maritime traffic data. Anomalies are detected as conflicts between vessel's potential in live data, and the local history-based potential field. The resulting detection performance is tested on AIS maritime tracking data from the Baltic region, and varies depending on the type of potential.

The potential field based approach contributes to maritime situational awareness and enables automatic detection. The results show that anomalous behaviours in maritime traffic can be detected using this method, with varying performance, necessitating further study.

Keywords: Anomaly Detection, Maritime Traffic, Potential Fields

1 Introduction and Motivation

Maritime traffic safety is of vital importance. According to the UN [Uni12], over 80 percent of the world trade traverses the seas by ship. Entities such as the Coast Guard continuously watch and safeguard the vessel traffic. Their work is aided by various surveillance technologies. Vessel maneuvers are primarily observed using marine radar. Today ships are usually equipped with more advanced navigational aids, such as an Automatic Identification System

(AIS) transponder. AIS is a surveillance system used on ships and by vessel traffic services to identify and locate vessels by electronically exchanging data with other nearby ships and fixed receiving stations. It integrates a positioning device (GPS), a gyrocompass, a speed meter and a rate of turn indicator to measure geographic location, course, heading, speed and rate of turn. Together with draught and static information (ID, name, size etc.) these measurements are packed in a standardized digital report and broadcast via VHF. The AIS communication range is further extended using satellites in low earth orbit for improved data exchange [Cha12].

The development of a globally available vessel tracking system opens the possibility of advancing maritime security far beyond simple local collision prevention. However, the few recent maritime detection solutions are still far from perfect [Riv09]. The existing tools are prone to false detections. Additionally, many maritime anomaly detection systems inform their users about detected anomalies but do not provide a meaningful explanation of why the detections were reported in the first place. The frequent misdetections and deficient explanations have a negative effect on users' trust in the detection system, further compromising its purpose.

This study introduces a novel method for data modelling and anomaly detection in maritime traffic. The novelty lies in employing the concept of potential field for data abstraction and representation. One of the aims is to improve understandability and maritime situational awareness, by visualizing the potential fields using modern rendering techniques. This would provide the maritime operators with a form of automated incident warning and analytical help in identifying traffic situations that merit further investigation.

The impetus for this study is that the content, quality and availability of the maritime traffic surveillance records have been drastically improved by the introduction of the international AIS standard in 2007. AIS transmission can be openly received and is standardised with regard to data content and format of messages. The use of AIS is currently mandated by an international maritime convention, however the quality of data cannot be fully relied upon. The precision of position and movement related data in AIS is limited by the quality of other integrated onboard sensors, such as GPS receivers or a compass. Moreover, some other static or voyage-related data, such as vessel name, size, destination port or ETA, are provided by the crew, thus making AIS prone to human errors, neglect and fraud. In practice, a large part of received AIS data is unusable due to misspellings, failure to update information and the like. For these reasons the AIS attributes considered unreliable are excluded from this investigation. This study limits the use of AIS attributes to the following set: longitude, latitude, speed, course, vessel type and timestamp.

2 Related Work

Two main types of approaches to maritime anomaly detection emerge: one focusing on defining anomalous behaviour explicitly, the other on inferring anomalous behaviour indirectly as an exception from the modelled normal behaviour.

Anomalous behaviour is often defined based on expert knowledge. This approach led to a number of studies identifying and listing anomalies [Lae09]. The collection of expert knowledge is conducted in various ways, such as expert surveys and workshops, practitioners brainstorming sessions [Lae09], or open maritime information extraction [Kaz13]. The resulting collection of expert knowledge is then used to define which vessel behaviours are considered anomalous. The definitions of known anomalous behaviours are then used as a base for anomaly recognition rules and detection.

This study may be counted among a number of approaches opposite to expert systems, in that they reverse the process of detection. These studies advocate that anomalous behaviour should not be defined directly, but implicitly — as a deviation from normal behaviour. Consequently, these approaches focus on the construction of a model of maritime traffic, representing all normal traffic behaviours. Tools used for defining normal behaviour stem from various scientific domains. Ristic et al. [Ris08] applied statistics to extract normal behaviour patterns from raw "messy" data. They define and model normal behaviour as motion inside areas implied by an extracted pattern, with normal speeds bound to them. The resulting "motion anomaly detection applied to AIS data" detected instances of anomalous trajectories (i.e. vessel passing through locations not belonging to the normal model) and velocity. Riveiro and Falkman [Riv09] proposed applying visualization techniques to enrich their rule-based anomaly detection and promote user interaction. Another joint work of Riveiro, Falkman and Ziemke [Riv08] combines a visual approach (self-organizing maps) with non-parametric statistics (density estimation by Gaussian mixture modelling) and probabilistic theory (Bayes theorem).

The need for automatic pattern extraction and detection has also been addressed by applying various machine learning techniques, e.g. neural networks [Per12], Similarity based Nearest Neighbour [Lax11], or proprietary solutions for normalcy learning [Rho06].

3 Method

The intention of this study is to develop a maritime data modelling method that enables extracting traffic patterns and detecting anomalies in a clear, understandable and informative way. Applying a potential field based method was inspired by game AI research, where it is used to create realistic bot movements. The potential fields used here to model maritime traffic are analogous to actual physical phenomenon of potential fields, e.g. electrostatic or gravitational [Jek81], and are described in a similar manner. The general idea in applying potential fields for maritime traffic is for the observed movement of each vessel to assign charges along its track. A collection of charges distributed over an area generates a potential field, which is locally weaker or stronger depending on the density and strength of surrounding charges.

The three main concepts derived from the physical potential fields are the charge accumulation, the decay of potential fields, and the distribution of potential around a charge.

3.1 Local Charge

The accumulation of charges is directly affected by the traffic surveillance data. Each vessel tracked by AIS is characterized by a collection of n numerical and textual properties. Those properties include vessel's static parameters, (e.g., name, flag, type), as well as the current state of its dynamic behaviour (e.g., speed, course, location), and are either inherently nominal or discretized to a nominal scale. A single vessel carries a set of charges of equal strength, representing its state and behaviour on these scales. For each AIS report, the set of charges c that a vessel carries is assigned to a location characterized by geographical position coordinates. Mathematically this can be expressed by a vector c_{lat_k,lon_l} with n components:

$$c_{lat_k,lon_l} = \left\langle c^1_{lat_k,lon_l}, c^2_{lat_k,lon_l}, ..., c^n_{lat_k,lon_l} \right\rangle, \tag{1}$$

where $c^1_{lat_k,lon_l}$ to $c^n_{lat_k,lon_l}$ are the component charges reflecting reported vessel properties: type, course, etc.; and lat_k, lon_l are the geographical latitude and longitude coordinates at point (k, l).

The total charge at a location is calculated as the sum of all local charges. In electrostatics the greater an electric charge is, the stronger the electric potential field that surrounds it. Analogically, the more vessels visits are reported at a location, the higher potential builds up in and around it. Hence the aggregate charge C_{lat_k,lon_l} accumulated at a location (k, l), over a time period τ would be computed as:

$$C_{lat_k,lon_l} = \sum_{t=0}^{\tau} c_{lat_k,lon_l} \tag{2}$$

3.2 Field Decay

This equation assumes no loss of charge. In continuous data collection over time that would allow charge to accumulate with no upper bound. This is undesirable, as it would undermine the ability to compare and follow trends of the maritime traffic behaviours over time. For example, once established real-world traffic patterns may get abandoned in time. It is desirable for the potential fields that model maritime traffic, to evolve over time to reflect such pattern changes. The addition of a field decay effect accomplishes this.

Researchers representing different approaches often address the problem of real time continuity by applying constructs such as a sliding time frame or a data window [Ris08; Bra10]. Potential field theory offers an alternative construct of potential decay. Adding a decay factor allows the charge at a location to be represented by a function of time:

$$C_{lat_k,lon_l}(t) = \sum_{t=0}^{\tau} d(t) c_{lat_k,lon_l} \tag{3}$$

where $d(t)$ is a non-increasing decay function with limit at zero. The function $d(t)$ describes the decrease of a local charge over time.

3.3 Field Distribution

Each local charge gives rise to a local potential. The potential field formed by a single charge c_{lat_k,lon_l} is most intensive in the location of the charge (k,l), and dissipates with increasing radius r. The dissipation of a physical potential field is represented by an equation specific to the type of field. Here it is defined as a decreasing function $f(r)$ of the distance r from the source charge. The distance r between points (k,l) and (x,y) is the Euclidean distance $\sqrt{(lat_k - lat_i)^2 + (a(lon_l - lon_j))^2}$, where a is the longitude coefficient with value in the range (0,1) compensating the disproportion between real geographical distances per unit of longitude versus latitude.

A global potential field is instantiated by geographically distributed local charges. The intensity of the field varies depending on the geographic location and is determined by the strength of the surrounding local charges affected by their decay, and the distance to them. Areas where a potential is very strong represent an emergent traffic pattern and describe a model of normal behaviour. Areas where a potential is very weak or non-existent signalize a lack of discernible normal traffic patterns.

An anomaly is here defined as a deviation from normal behaviour, thus an observed vessel behaviour that does not conform to the normal model described by the potential fields, is considered anomalous. This is made feasible by the fact that the vast majority of maritime traffic occurs normally, i.e. in wide understanding not abnormal, and can be described by such a model of normal behaviour. Based on this assumption, the presence of potential represents normal behaviour, and its absence — an anomaly.

One of the advantages of the proposed method is the ability to detect and signalize different severity of perceived threats (i.e. anomalies), depending on the intensity of the discrepancy with the potential fields. In this study the anomaly levels are determined using minimal potential thresholds. Another benefit is the possibility of visualizing the potential fields for enabling prompt perception and comprehension of what exactly the violation of the normal models entails.

4 Potential Fields Applied

For the purpose of this study, a maritime traffic modelling and detection system named STRAND (*S*eafaring *TR*ansport *AN*omaly *D*etection) was prototyped. It implements the proposed method of potential fields applied to maritime surveillance data.

4.1 Discretization

In AIS-based maritime surveillance, the continuous parameters are time, space, speed, course and other related. For the potential field to generalise knowledge of normal behaviour (and for computational efficiency reasons) the continuous parameters need to be binned. This section describes how the variables are discretized.

The time is represented by a POSIX timestamp, which limits the precision to 1s. Each AIS message contains such a timestamp, however the global load of AIS messages is updated once every 90s, making 90s the basic the time unit for detection. Latitude and longitude precision in AIS transmission (0.0001') translates to ca. 18.5cm in the real world. Meanwhile, the declared maximal precision of navigation devices in vast majority of AIS transponders is 10m [IMO02]. For data modelling and detection this precision is adjusted to represent an approximately square grid with tile size 10m. Course and speed over ground are stored without alterations with a precision of 0.1° and 0.1 knot respectively. For maritime traffic modelling and detection, however, they are binned into ranges. Course is divided into 8 equal intervals: *N*, *NE*, *E*, *SE*, *S*, *SW*, *W*, and *NW*. Speed ranges are not equal in size, and correspond to the speed classes common in maritime circles, from *Static* (0–1 knot) to *Probably flying* (exceeding 60 knots). The precision of course and speed is reduced to a nominal scale for twofold reasons: to build an understandable parameter selection in the user interface, and to define the speed and course value granularity for data modelling and detection sensitivity.

The other attributes are the nominal vessel type (*Passenger*, *Cargo*, *Tanker*, etc.) and the time of day. The time of day divides 24 hours into four equal time bins: *Morning* (6–12), *Afternoon* (12–18), *Evening* (18–24), and *Night* (0–6).

4.2 System Implementation

The prototype system implements the three aspects of the proposed method. Live AIS tracking data is gathered over time in order to collect a representative traffic history required for building a normal model. The first element of the method concerning the local charge is implemented as the sum over the iterated traffic history. A single iteration handles one stored AIS position report consisting of the attributes: latitude, longitude, course, speed, vessel type and timestamp. The report contributes to the cumulative charge of the nearest grid node represented by the discretized latitude and longitude pair. The complete grid of local charges is stored once for every 24 hours.

The field decay is implemented as an exponential decrease of the charge. The prototype builds a normal model based on a real world AIS data set spanning 20 days. The aggregated charge is divided by a constant in each iteration over the days of the traffic history, making the decay function $d(t)$ from equation 3 an exponential function of time. For testing purposes a daily charge decay rate of 10% was used. In large or continuous data sets there is a need for a termination condition, excluding most decayed charges from the normal model.

Field distribution was implemented using the two-dimensional Gaussian smoothing equation [Jek81]. The local potential value is evaluated as:

$$P_{lat_k,lon_l}(t) = \sum_i \sum_j \frac{1}{2\pi\sigma^2} e^{-\frac{(lat_k - lat_i)^2 + (lon_l - lon_j)^2}{2\sigma^2}}. \quad (4)$$

The standard deviation is set to one grid side length, i.e. approximately 10m. The radius of the smoothing around location (k, l) is defined by the latitude and longitude limits.

An observed consequence of using AIS position reports as charges is an imbalance between the amount of charge generated by vessels traveling with different speeds. The need for compensating this issue in the modelling and detection phase was addressed by introducing a charge multiplier. All components of a single charge set, representing one AIS message, are multiplied by the square root of the vessel speed, thus moderately weakening the charges dropped by very slow vessels and increasing the charge for faster ships.

A Web-based prototype system was built for this study. It was implemented using the Django Python Web framework. Figure 1 presents a view of the system interface with an example of map-based display of a potential field and detection. Small black arrows represent vessels conforming to the normal behaviour patterns. The anomalously behaving vessels are marked by larger red arrows.

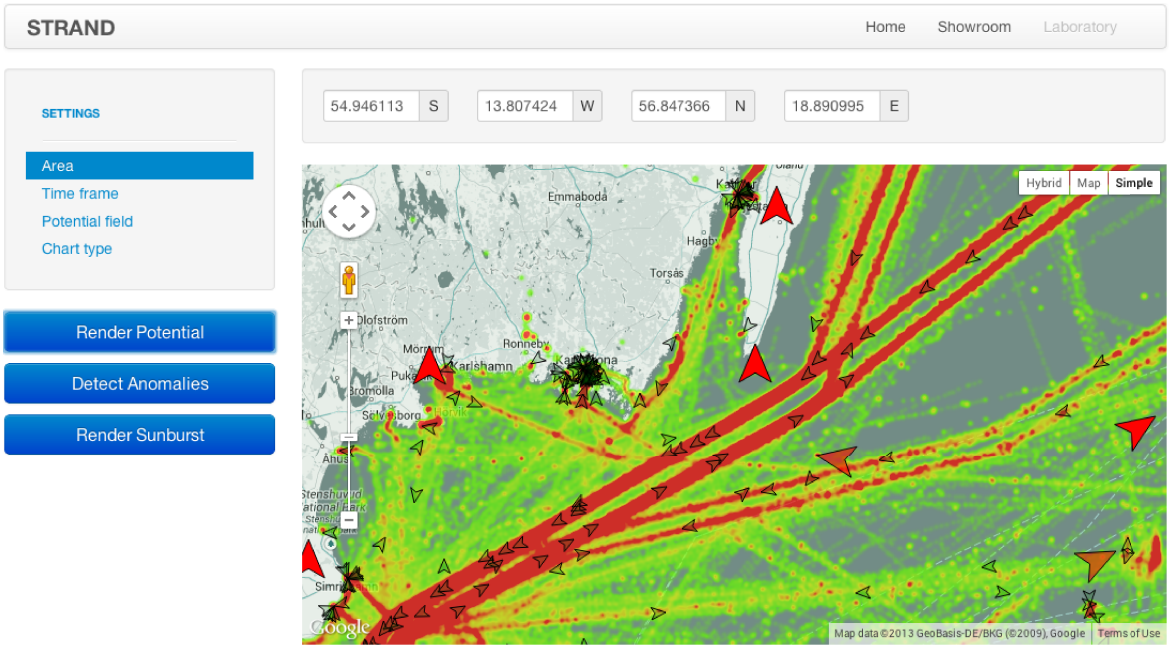


Figure 1: STRAND user interface with area, time frame and potential type selection.

The STRAND user interface includes a map with overlays (from Google Maps API v3), a set of controls and a page navigation panel. The controls include menus for setting the coordinate limits, potential metric (speed, course, vessel type, day time) and range of the metric value, as well as the optional time frame.

5 STRAND System Case Based Demonstration

In the detection process, the examined position reports are crosschecked with all potential field values for the corresponding position grid coordinates. The current version of STRAND detects anomalous incidents associated with faulty position, time of day, speed, course and type of the vessel. An example of a detection situation is presented in Figures 2, 3 and 4.

In an observation of traffic nearby the Swedish Blekinge coast most vessel behaviours are considered normal (small black arrows). A red arrow icon visible on the right side of all

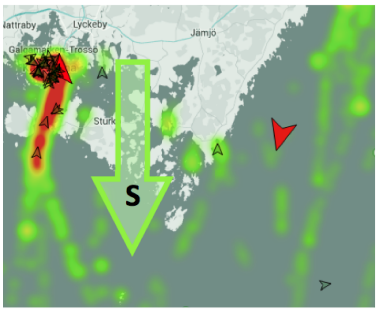


Figure 2: Course S.

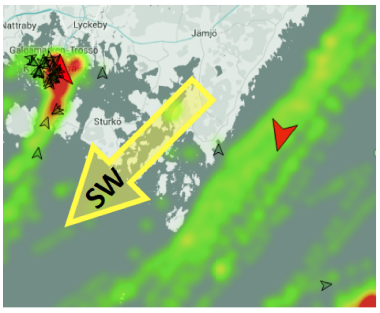


Figure 3: Course SW.

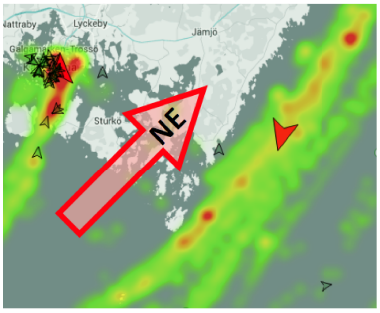


Figure 4: Course NE.

images marks an anomaly. This situation is illustrated in the figures 2, 3 and 4, with relevant potential fields for courses S, SW and NE respectively. The potential intensity is represented by heat maps overlaying the geographical map, i.e. the colour palette from green to red represents increasing potential values. In the observed situation, the anomalous vessel is moving southwards in an area where southward traffic is unusual (Fig. 2) and dominant traffic course is northeast (Fig. 4). The southward traffic patterns (Fig. 2) are located further to the west and to the east from the vessel's position, but absent at the current position. In figure 3 the observed southwestern traffic pattern is weak, but present at vessel's location.

Based on such observations, it is possible to notice and understand the nature of the anomaly and, furthermore, to formulate a recommendation how to correct the anomalous behaviour. In this case one would recommend correcting the course to SW to conform to the local SW traffic pattern (Fig. 3), or, if the southern course is strongly desired, correcting position and route to follow one of the southern traffic patterns (Fig. 2).

The detection itself is performed only based on the potential fields directly corresponding to the vessel behaviour. In the observed case the comparison of vessel course to the southward traffic pattern raised the alarm. The NE and SW traffic patterns were used to enhance comprehension and situational awareness, however they did not participate in the computation that led to that particular detection.

6 Discussion

The presented tool demonstrates modelling and visualizations of attributes: course, speed, daytime and vessel type, using three concepts; the charge aggregation, the distribution of potential field around its source, and the field decay. Implementation of the method brings to light the benefits of using a normal model in contrast to expert based anomaly recognition rules. The normal model is built based on actual maritime traffic avoiding possible human bias and limits of knowledge or comprehension of behaviours occurring in maritime traffic.

By the implementation of the STRAND prototype a number of issues were raised and

described but not exhaustively addressed. The system demonstrates the applicability of the method using a set of prototypical parameters and algorithms, which are not intended as the ultimate solution. The numerical parameters that require tuning are e.g., the grid size, and the detection thresholds for speed, course, vessel type and time of day. Algorithmic issues include the total charge calculation, the potential field dissipation and decay equations, and the speed compensation.

The grid size tuning is especially challenging since for one it represents a non-flat area — an ideal grid would adjust the longitude to always fit the same Euclidean distance. Secondly, the grid is uniform, the node size for open sea and harbour-like areas is the same. The precision of AIS data gives the possibility to vary the density of geographical grid. This allows for fine-tuning a grid size starting from figuratively the size of the ship to thousands of meters. A vessel with a speed of ten knots will travel almost 500 meters within 90 seconds, i.e. the possible discretization of involved values is far too precise to be handled by the prototype tool. Therefore a balance where each stored vessel position should correlate to a unique grid position without being too spread out over the grid net is desirable. In practice this can be achieved by introducing different local grid precisions, i.e. sizes of nodes in the potential field, depending on the local maritime situation (e.g. small grid size in harbours, larger on the open sea). In the current STRAND tool this is not implemented.

7 Conclusion and Future Work

The STRAND prototype system demonstrates the applicability of the proposed method. The three aspects of potential field theory: charge accumulation, potential decay and dissipation, enable the modelling of vessel traffic. The resulting normal model facilitates customizable visualization and anomaly detection. An advantage of the method is the ability of creating a normal traffic model based on the traffic history, without a need for expert knowledge. The geographical map-based grid is filled by a potential field derived from the observed traffic. Anomalies are identified as a lack of normal behaviour — local absence of potential.

The outline implementation structure provided by the method opens space for algorithmic and computational optimization. Field decay and dissipation functions used in this study, exemplified and enabled the demonstration of the devised method, but also raised questions for future study. Another open issue is the visual anomaly reporting as well as normal model representation. Here choices need to be made concerning the manner of displaying extracted patterns and traffic information in the user interface. The discretization of AIS attributes and its impact on detection performance should also be studied further.

References

[Bra10] C. BRAX, A. KARLSSON, S. ANDLER, R. JOHANSSON, and L. NIKLASSON: “Evaluating precise and imprecise State-Based Anomaly detectors for maritime surveillance”. In: *Proc. of the 13th Conf. on Information Fusion (FUSION)*. 2010.

[Cha12] R. CHALLAMEL, T. CALMETTES, and C. N. GIGOT: "A European hybrid high performance Satellite-AIS system". In: *Proc. of the 6th Advanced Satellite Multimedia Systems Conf. (ASMS) and 12th Signal Processing for Space Communications Workshop (SPSC)*. 2012.

[IMO02] *Guidelines for the onboard operational use of shipborne automatic identification systems*. International Maritime Organization, 2002.

[Jek81] C. JEKELI: "Alternative methods to smooth the Earth's gravity field". In: *Reports of the Department of Geodetic Science and Surveying, Report 327*. Ohio State University, Columbus, 1981.

[Kaz13] S. KAZEMI, S. ABGHARI, N. LAVESSON, H. JOHNSON, and P. RYMAN: "Open data for anomaly detection in maritime surveillance". In: *Expert Systems with Applications* 40.14 (2013), pp. 5719–5729. ISSN: 09574174.

[Lae09] J. van LAERE and M. NILSSON: "Evaluation of a workshop to capture knowledge from subject matter experts in maritime surveillance". In: *Proc. of the 12th Int. Conf. on Information Fusion*. 2009.

[Lax11] R. LAXHAMMAR and G. FALKMAN: "Sequential Conformal Anomaly Detection in trajectories based on Hausdorff distance". In: *Proc. of the 14th Int. Conf. on Information Fusion (FUSION)*. 2011.

[Per12] L. P. PERERA and P. OLIVEIRA: "Maritime Traffic Monitoring Based on Vessel Detection, Tracking, State Estimation, and Trajectory Prediction". In: *IEEE Transactions on Intelligent Transportation Systems* 13.3 (2012), pp. 1188–1200.

[Rho06] B. J. RHODES, N. A. BOMBERGER, M. SEIBERT, and A. M. WAXMAN: "SeeCoast: Automated Port Scene Understanding Facilitated by Normalcy Learning". In: *Proc. of the Military Communications Conf.* 2006.

[Ris08] B. RISTIC, B. L. SCALA, M. MORELANDE, and N. GORDON: "Statistical Analysis of Motion Patterns in AIS Data: Anomaly Detection and Motion Prediction". In: *Proc. of the 11th Int. Conf. on Information Fusion*. 2008.

[Riv08] M. RIVEIRO, G. FALKMAN, and T. ZIEMKE: "Visual Analytics for the Detection of Anomalous Maritime Behavior". In: *Proc. of the 12th Int. Conf. Information Visualisation*. 2008.

[Riv09] M. RIVEIRO and G. FALKMAN: "Interactive Visualization of Normal Behavioral Models and Expert Rules for Maritime Anomaly Detection". In: *Proc. of the Sixth Int. Conf. on Computer Graphics, Imaging and Visualization*. 2009.

[Uni12] *World Economic Situation and Prospects 2012*. United Nations, 2012.

Corresponding author: Ewa Osekowska, Blekinge Institute of Technology, School of Computing, 37179 Karlskrona, Sweden, phone: +46 455 385 830, e-mail: ewa.osekowska@bth.se

Trajectory Optimisation for a Manoeuvre Guidance System in Inland Water Traffic

Alexander Born, Iván Herrera-Pinzón
German Aerospace Center (DLR)

Abstract

The increasing traffic on inland waterways demands the adaption of extended routing and manoeuvre planning in order to increase the safety and to improve the time and resource efficient transport of goods and passengers. A basis is the calculation of an optimal trajectory for a vessel taking infrastructures in the fairway into account. A further prerequisite is the ability to monitor and to react to changes in the fairway, such as other traffic participants, and to update the trajectory in a timely manner. Mathematical optimisation methods can be used to meet these requirements. This paper discusses two different algorithms and compares both in terms of calculation effort and real-time computation. Finally one method, the Optimal Control Problem-solver (OCP), will be chosen to be implemented in the manoeuvre guidance system.

Keywords: Trajectory Optimisation, PNT-Unit, Maneuver Guidance System

1 Introduction

The seaborne and inland water transports of goods are an essential basis for Europe's economic development, competitiveness, and prosperity. Inland water transportation becomes more important taking into account current ecological and economic challenges. The inland water traffic system has also a direct impact on the quality of life of citizens, both as tourists and inhabitants of islands and peripheral regions.

The need to enhance the maritime traffic system was recognised by the International Maritime Organization (IMO) and resulted into the initiative 'e-Navigation' initiative initiated in 2006 [IMO07]. Europe started similar activities in the inland water sector with the development and implementation of International River Information Systems (IRIS) to provide traffic and transport related information [IRI13]. External traffic and transport information in combination with vessel specific state information are the necessary data basis to enable a time- and resource-efficient manoeuvring of vessels.

The project Precise and Integer Localisation and Navigation in Rail and Inland Water Traffic (PiloNav), whose goals are described in Chapter 3, meets this challenge for inland water transport. In this context the passing of bridges, the locking of ships and docking manoeuvres in ports are crucial phases with a need for optimised vessel trajectories.

The remainder of the paper is structured as follows. Where Chapter 2 gives a short overview over the related work, Chapter 3.1 discusses the challenge to provide highly accurate Position, Navigation and Timing (PNT) data of the own vessel as input for safety-critical applications, such as a manoeuvre guidance system. Chapter 3.2 introduces the modular concept of an inland water manoeuvre guidance system. Before concluding the paper in Chapter 5, two different mathematical optimisation procedures will be analysed in Chapter 4, using simulated and data resulting from a measurement campaign.

2 Related Work

The topic of a real-time trajectory optimisation implementation has been extensively discussed by Miele et al in [Mie05; Mie06; Mie99; Tze98] with the provision of the so-called Multiple-Subarc Gradient Restoration Algorithm via the usage of realistic vessel's kinematic models [Mie03; Mie74] plus complex -though effective- cost functional, demonstrating its suitability for maritime applications in open waters by virtue of its high performance and accuracy, lacking however of the use of autonomous sensors for the enhancement of the calculation of the vessel's position and navigational state.

With his work on the direct solution of optimal control problems via the sequential quadratic programming approach in [Fab12; Fab98; Fab08a; Fab08b; Fab13], Fabien has developed a robust set of efficient programming tools able to handle the trajectory optimisation problem allowing the definition of several functional and kinematic models dependent of user-defined variables, with an interesting susceptibility for the real-time applications.

In his PhD work [Lut11], A. Lutz goes further and proposes a collision detection system for inland waters with the use of modern technologies, such as Global Navigation Satellite Systems (GNSS), Automated Identification System (AIS) and other ship-side sensors, incorporating the use of hydrodynamic models to increase to adequacy of his models but missing the usage of integrity functions to determine and to monitor the quality and reliability of the calculated PNT-data, which is a must considering troubling scenarios and complex manoeuvres.

3 Project PiloNav

The goal of PiloNav is the development of a generic location platform which can be used on different transport carrier to determine highly accurate and integer position, navigation and timing data with the focus on rail and inland-water traffic.

Due to its availability Global Navigation Satellite Systems are used in every traffic carrier. When it comes to signal degraded environments, the requirements, formulated exemplary for

maritime and inland water traffic in Table 1, cannot be achieved using GNSS only [Vie12b]. Therefore, carrier-specific sensors (IMU, radar, optical sensors, etc.) will be merged with position, navigation and timing information obtained by GNSS and therefore to form an integrated navigation system (INS). In case of inland-water traffic this is referred to as a Positioning, Navigation and Timing-Unit (PNT-Unit) and in case of rail traffic as a Train Location Unit (TLU) respectively.

On the application layer this PNT data will be used as input for driver assistance and manoeuvre guidance systems which continuously provide reliable information and assist the user with critical manoeuvres in order to optimise rail and inland water traffic and to meet the requirements in terms of efficiency and environmental challenges [Vie12b; Alb13]. This work, however, focuses on the inland water aspect, where the inland water transport-relevant goals of the project can be formulated as:

- the development of an inland water PNT-Unit,
- the development of a manoeuvre guidance system to enable a time- and resource efficient passing of bridges, locking of vessels, and to plan and to perform evasion manoeuvres in case of oncoming traffic,
- and the validation of the integrated manoeuvre guidance system with simulated data and by experimentation.

The next paragraph shortly describes the components of the PNT-Unit as data source.

3.1 Concept of the inland water PNT-Unit

In order to evaluate the determined PNT-data, accuracy and integrity requirements have to be formulated. Requirements on navigational parameters are defined within the mentioned 'e-Navigation' initiative by the IMO for seaborne navigational systems and services [IMO07; IMO07a; IMO09], where inland water is defined as sections between the open sea and the harbour. However, in this work denotes inland water application as vessel navigation on rivers. First requirements on navigational parameters are formulated by the projects IRIS Europe 2 and PiloNav [IRI13; Vie12b]. Table 1 gives an overview over the requirements on the PNT data for inland water traffic and compares them with requirements formulated by the IMO.

Considering the opportunity of trajectory optimisation, the projects Iris Europe II and PiloNav formulated extended requirements on navigational parameters for aimed project developments which are significantly higher than maritime requirements formulated by the IMO.

Due to the good availability of global navigation satellite systems, GNSS became a widely used technique for positioning and navigation on inland water vessels. They are also part of the maritime communication system AIS (Automated identification System), which is also widely used in inland water traffic. Here, only 1 frequency receivers are mainly used to

Table 1: Accuracy requirements on PNT data for Inland water traffic [IMO09; Vie12b; IRI13]

3-D Pos. Req.	IMO	PiloNav/IRIS Europe 2
Horizontal Position Accuracy	10.0 m (Regular Waterways)	3.0 m (Regular Waterways)
	1.0 m (Port Approach)	0.10 m (Locking)
	0.1 m (Automated Docking)	0.10 m (Bridge Passing)
Vertical Position Accuracy	$1.0 - 10.0$ m (Regular Waterways)	0.10 m (Bridge Passing)
Heading	0.75°	0.1° (Locking)
Integrity Risk	$10^{-5}/15$ min	$10^{-5}/15$ min

determine the position and attitude of a vessel for information purposes. Using this technique a determination of the navigational state with the desired accuracy is not possible [Lan13]. Consequently, two GNSS sensors are used as main sensor for the generic location platform. Compared to deep sea, inland water vessel navigation has the advantage that ground-based reference services can be used to improve the positioning by satellite based sensors. Therefore, Real-Time Kinematics (RTK) will be used to complete the satellite-based part of sensors.

Measurement campaigns show that the application of GNSS-based sensors and additional GNSS-based services only does not suffice the demanded requirements due to environmental or infrastructural conditions as shown in [Vie12b]. Moreover, additional sensors have to be integrated to compensate loss of satellite signals and multipath effects in signal degraded environments. In particular, an Inertial Measurement Unit (IMU) will be used to measure accelerations in all axis as well as rotation angles. This type of measurement can be used to track and to predict the movement of the vessel over a short period.

3.2 PNT-Unit based Vessel Manoeuvre Guidance System

Knowing the position and navigational states of a vessel only is not sufficient to design and conduct sensible vessel movements. Uncertainties which have an impact on the traffic flow have to be detected and analysed. These uncertainties one distinguishes between static (bridges, locks, quay walls) and dynamic (position and navigational states of other vessels) factors. Both have to be analysed to be able to predict the development of the traffic situation and respond in a timely manner. While digital maps, such as Electronic Navigational Charts (ENC), can be used to recognise the occurrence of static factors affecting the calculation of optimal trajectories or evasive manoeuvres, dynamic factors have to be monitored using ship-borne or so-called application sensors such as radar and the communication system AIS. Figure 1 displays the architecture of the manoeuvre guidance system for inland water vessels. The Sensor & Data module delivers further PNT-independent information used for the manoeuvre guidance system. This is information about the vessel itself (size, hull geometry, thresholds for manoeuvring data such as max. velocity and turn rates) and external information (AIS-data of other vessels, radar images or digital maps like the Electronic Navigational Chart (ENC)). Influences such as wind, currents (natural and self-induced due to interaction with river banks, locks and walls) or the Squat effect will not be taken into account as this

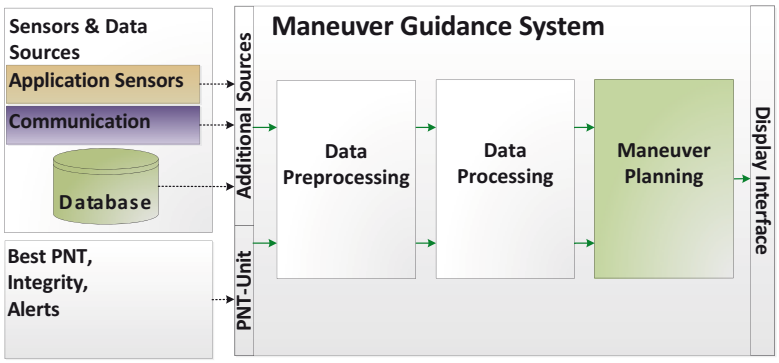


Figure 1: Architecture of the PiloNav Manoeuvre Guidance System for Inland Water Vessels

system is implemented as a first application defining requirements for using integer PNT data as derived by the PNT-Unit described in the previous section.

The Data Preprocessing contains the determination of the traffic situation and prediction (Tactical Traffic Images - TTI), the positioning of the ships contour and the consideration of water level in static environmental data.

Within the Data Processing module distances to obstacles will be determined and a short-term collision detection performed. For a detailed description of these modules the reader is referred to Vierhaus et al in [Vie12a].

This information is used as data input for the manoeuvre planning module. The subject of this work, however, is the computation of the elements for the trajectory optimisation which is part of the manoeuvre planning module (green box in Figure 1).

Based on the PNT-data and TTI the manoeuvre guidance system provides the elements of the optimal trajectory consisting of velocities and heading information as well as strategic navigational tasks such as encounters with other ships as well as tactical tasks like bridge and lock passing. In the first step this allows the manoeuvre guidance not only to follow the defined trajectory and returning in case of deviation. It is also capable to interact with other vessels, i.e. to plan evasion or overtaking manoeuvres, and providing course and velocity for the upcoming stretch of the river to the skipper in a timely manner. An interaction in terms of bidirectional planning of theses manoeuvres with other vessels will not be implemented as this exceeds the AIS specification. The planned manoeuvres will be optimised in terms of the criteria safety as well as time- and resource-efficiency. In the next chapter two different mathematical approaches are be presented and compared.

4 Trajectory Optimisation for a Vessel Manoeuvre Guidance System

Once the PNT-data have been collected and properly processed by the manoeuvre guidance system, the topic of collision avoidance becomes the central issue for inland water navigation; thus, inland vessels should include an optimal manoeuvring system capable to address these

scenarios. Over time, several physical models have been developed to work out this kind of manoeuvres [Mie05; Mie06; Mie99; Tze98; Yav97; Vie12a], but it is widely accepted that in order to cope with the trajectory optimisation procedure, these models should meet at least the key criterion of minimising the course deviation with smallest control effort [Mie05; Mie99].

Based on these ideas, in order to calculate optimised trajectories for the PiloNav test vessel, the following simplified model describing the kinematics of a ship is implemented: For navigation in river corridors, it is considered a model where the vessel moves only in the $\mathbf{x}-\mathbf{y}$ plane with the orientation defined by the Course over the Ground ($\mathbf{CoG}$) - relative to the $\mathbf{x}$ axis. The speed of the ship is defined by its velocity $\mathbf{v}$ in the direction of the $\mathbf{CoG}$, with the acceleration $\mathbf{a}$ also pointing towards this direction. The angular velocity of the ship is regarded as the so-called **Rate of Turn** ($\mathbf{r}$), $\dot{\mathbf{CoG}} = \mathbf{r}$. Therefore, within the scope of this problem, $\mathbf{a}$ and $\mathbf{r}$ are considered the control inputs. Hence, the optimal control problem can be written as:

$$\begin{aligned} \text{State Variables:} \quad y &= [x, y, v(t), CoG(t)]' \\ \text{Control Variables:} \quad u &= [a(t), r(t)]' \end{aligned} \tag{1}$$

$$\text{Cost Functional:} \min_{t_f, a(t), r(t)} \left[c_1 \cdot t_f + \int_{t_0}^{t_f} \left(c_2 \cdot a^2(t) + c_3 \cdot r^2(t) \right) dt \right] \tag{2}$$

Subject to

$$\begin{aligned} \text{Dynamic Equations:} \quad & \dot{y} = [v(t)\cos(CoG(t)), v(t)\sin(CoG(t)), a(t), r(t)]' \\ \text{Inequality Constraints:} \quad & [(0, v_u), (a_l, a_u), (r_l, r_u)]' \\ \text{Path Constraints:} \quad & d(\mathbf{x}, \mathbf{Obstacle}) \geq \varepsilon \end{aligned} \tag{3}$$

Where the sub-indices l and u denote the lower and upper values, respectively. Finally, the inclusion of static and moving elements that alter the original course of the ship are taken into account by mean of a set of path constraints, where the euclidean distance d between the current position of the ship and the location of the obstacle is calculated. Both state and control variables are bounded by the inequality constraints, determined for the PiloNav demonstration vessel in previous measurement campaigns as the average of the observed values for each variable, and defined as:

$$\begin{aligned} 0\ [m/s] \leq \quad & v \quad \leq 5.5\ [m/s], \\ -0.1[m/s^2] \leq \quad & a \quad \leq 0.05\ [m/s^2], \\ -120^\circ/min \leq \quad & \mathbf{r} \quad \leq 120^\circ/min. \end{aligned} \tag{4}$$

While is true that the height component plays an important role in all navigation applications, given the nature of the used data used for the scope of this work where the majority of

the measured points have vertical clearance, it is sufficient to consider a 2D optimisation approach; however, since the safe passing under bridges can be eventually threatened by the height clearance, further models involving 3D optimisation shall be considered.

4.1 Available Tools

For evaluating the suitability of the aforementioned kinematic model, several scenarios under different traffic conditions are simulated and processed to obtain the desired optimal paths and, in order to ensure the quality of these trajectories, two different software packages where tested. Among the variety of software packages available for the solution of optimal control problems, these two were selected due to their versatility when recreating several kind of scenarios with the key feature of the multiple-phase definition of the problem, in the pursuit of a suitable tool able to deal with PNT-data in near real-time. A brief description of these software packages is presented below.

MatLab Optimal Control Software: GPOPS-II

The *Gauss Pseudospectral Optimisation Software* (GPOPS-II) is a set of MATLAB routines intended to solve optimal control problems. GPOPS-II uses the Gauss pseudospectral method developed at The University of Florida to transform the optimal control problem into a nonlinear programming problem (NLP) [Rao10], which are solved using the large-scale NLP solver IPOPT: the Interior Point OPTimizer [Wac06].

OCP: An Optimal Control Problem Solver

The package *OCP: An optimal control problem solver* (OCP), is a group of C-code routines developed by B.C. Fabien at the University of Washington. OCP is a multi-purpose tool which solves the optimal control problem by transforming them to a nonlinear programming problem, which are solved using the sequential quadratic programming (SQP) technique [Fab13].

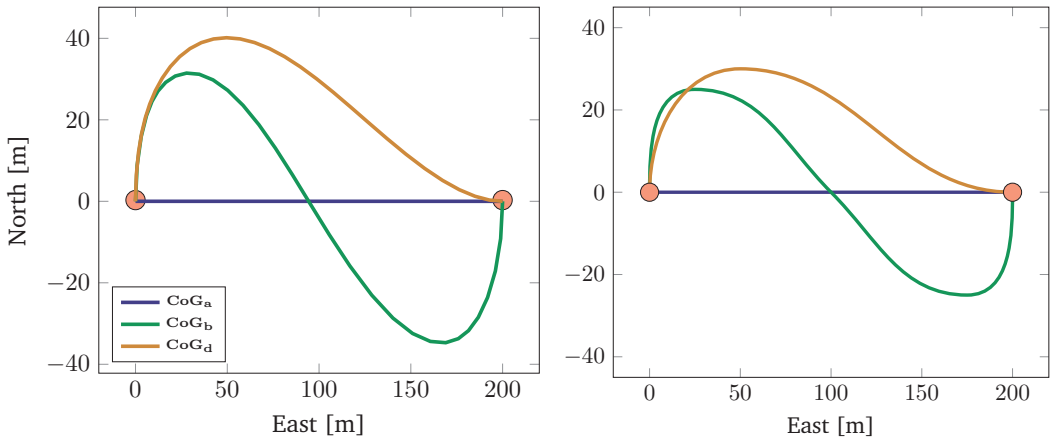
4.2 Numerical Simulations

Further on in this document three scenarios for the path optimisation with different traffic conditions will be considered: Way point approaching, way point approaching while a stationary obstacle has to be avoided and way point approaching while avoiding an oncoming vessel; which, according to previous measurement campaigns, are the most frequent manoeuvres carried out when sailing inland river corridors. Of particular interest for the scope of this work is the performance of the the software tools against real-time data, thus a quantitative analysis of their efficiency will be regarded.

Way point Approaching

With no elements obstructing the path of the vessel, this scenario attempts to solve the simple task of approaching the following destination point. It is required to find the optimal trajectory that will move the vessel from $[0\ m, 0\ m, 2.5\ m/s, \mathbf{CoG}_0]'$ to $[200\ m, 0\ m, 2.5\ m/s, \mathbf{CoG}_f]'$ minimising time, acceleration and rudder movements.

While is obvious that, when considering a planar coordinate system, the optimal path between the two points is a straight line; it is important to note that when the $\mathbf{CoG}$ is altered the trajectory will display a more complex behaviour. Figure 2 shows the trajectories calculated with the two different packages for three different combinations of $[\mathbf{CoG}_0, \mathbf{CoG}_f]$, namely $[0, 0]$, $[\pi/2, 0]$ and $[\pi/2, \pi/2]$, and labelled as $\mathbf{CoG_a}$, $\mathbf{CoG_b}$ and $\mathbf{CoG_c}$, respectively. To illustrate the performance of the different software packages, Table 2 shows the terminal



(a) OCP **(b)** GPOPS-II

Figure 2: Way point approaching

time (time needed for the ship to complete the manoeuvre), number of iterations and the time of calculation employed by the software to solve each one of the aforementioned cases. An

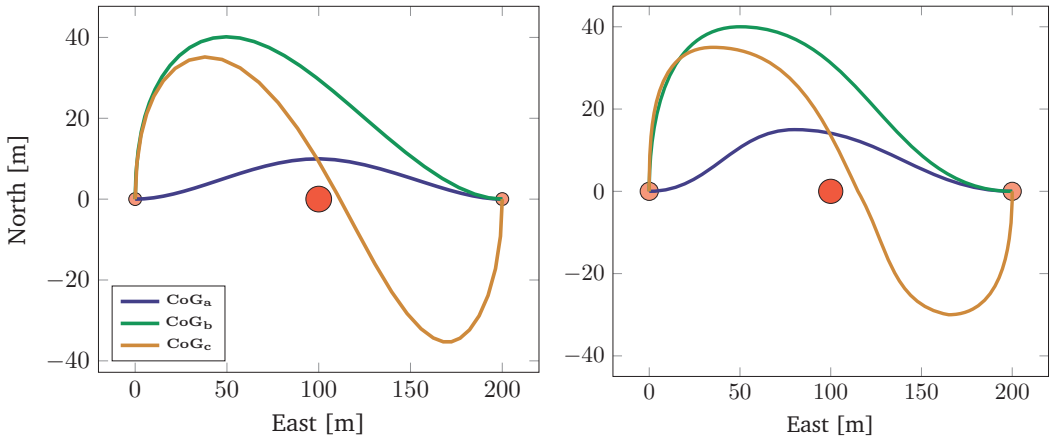
Table 2: Parameters of the solutions (I – $[0, 0]$, II – $[\pi/2, 0]$, III – $[\pi/2, \pi/2]$)

$[\mathbf{CoG_0}, \mathbf{CoG_f}]$	OCP			GPOPS-II		
	I	**II**	**III**	**I**	**II**	**III**
Terminal time [s]	57.8	81.8	115.2	59.7	79.4	106.8
Solution time [s]	0.17	1.1	1.3	2.67	4.15	7.40
No. of iterations	46	74	161	83	142	237

evident qualitative and quantitative difference, yet in both solutions satisfactory, is observed in both Figure 2 and Table 2, which is supported by the fact that both tools use different methods to address the solution of the underlying NLP.

Static Obstacle Avoidance

A stationary obstacle located at the position $(100\ m, 0)$ has to be avoided. Within this scenario a buffer of $10\ m$ around the obstacle has been defined as safely margin. To meet this requirement the path constraint $\sqrt{(x-100)^2+(y-0)^2} \geq 10$ should be considered in the model. Figure 3 illustrates the manoeuvres for the same scenarios described on Section 4.2. Clearly the incorporation of obstacles within the model define expressly the results, not only



(a) OCP **(b)** GPOPS-II

Figure 3: Static Obstacle Avoidance

by affecting the trajectory and the terminal time of the manoeuvre, as can be seen in Table 3, but also by compromising the times of calculation and the number of iterations used to solve numerically the problem. Once again both pieces of software show disparate solutions; the

Table 3: Parameters of the solutions (I – $[0, 0]$, II – $[\pi/2, 0]$, III – $[\pi/2, \pi/2]$)

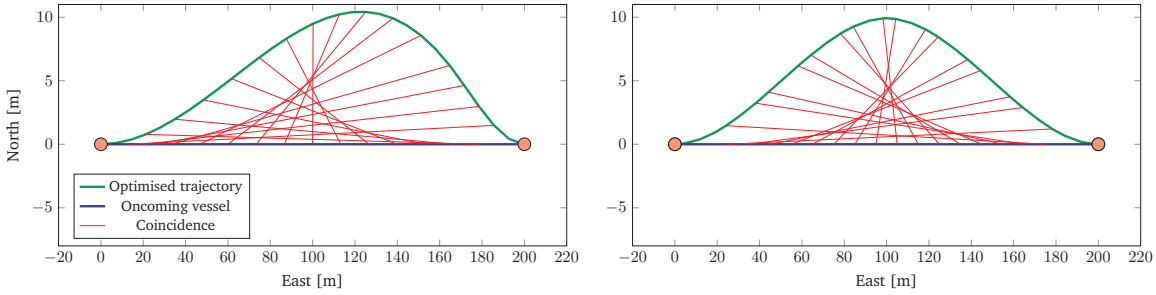
$[\mathbf{CoG_0}, \mathbf{CoG_f}]$	**Static Obstacle**						**Oncoming Vessel**	
	OCP			**GPOPS-II**			**OCP**	**GPOPS-II**
	I	**II**	**III**	**I**	**II**	**III**		
Terminal time [s]	58.0	81.8	115.4	60.1	85.4	106.3	90.4	99.7
Solution time [s]	0.44	0.48	0.79	3.58	4.15	10.4	2.23	3.56
No. of iterations	44	63	115	123	162	253	137	179

determined trajectories display similar behaviour, although both solutions match much less than those in Section 4.2, but specially the technical details of the solutions exhibit a bigger contrast that the ones presented above. While both solution are satisfactory in the sense that they are feasible in terms of manoeuvring, the technical parameters of the one obtained by using the OCP tool proof its usability for this kind of applications.

Oncoming Vessel Avoidance

An overtaking manoeuvre involving an oncoming vessel with constant velocity ($4\ m/s$) and course over the ground (π) is considered. The safety region for the avoiding vessel is defined by a buffer of $10\ m$. Therefore, it is sought the optimal trajectory that will take the vessel state from $[0\ m, 0\ m, 4\ m/s, 0]'$ to $[200\ m, 0\ m, 4\ m/s, 0]'$. A path constraint of the form $\sqrt{(x-(100-4\cdot t))^2+(y-0)^2} \geq 10$ should be then included in the model.

Figure 4 and Table 3 illustrate the results obtained and exemplifies again the different -both valid- solutions that can be achieved due to the conceptual differences that were implemented in each package. It is worth to mention that, in spite of they seem very different, the



(a) OCP **(b)** GPOPS-II

Figure 4: Oncoming Vessel Avoidance

optimised trajectories elude the oncoming ship while maintaining the desired safe distance and minimising time, acceleration and rudder movements. Although both trajectories are equally feasible in terms of manoeuvring and both accomplish the optimisation task, the OCP solution shows again better results in terms of computational effort which suggests a better performance when working with real-time data.

4.3 Approach for PNT-Data based Optimal Trajectories

After an evaluation of theoretical results that extend those scenarios discussed in the last section, for convenience in the implementation and the latency of the results, it was decided to assess the performance of the tool OCP, mentioned in Section 4.1, with real PNT-Data collected in a measurement campaign on the river Mosel in the city of Koblenz (Germany) in August 2012. A set of ca. 400 points collected and processed by the PNT-Unit were used to test the robustness of the software in a real environment over the simple case of way point approaching (see Table 4). To overcome need of actual way points for the passage of the vessel at every stage of the channel, a set of coordinates representing the ideal trajectory of a ship within the river was provided by the German Federal Waterways and Shipping Administration for Navigation Techniques.

The ideal trajectory, calculated as the average of a series real trajectories corresponding to different kind of vessels transiting the river for a period of one week that includes the day of the measurement campaign, was considered as the best practice scenario for the navigation

Table 4: Sample of Koblenz Data

Timetag	East (UTM) [m]	North (UTM) [m]	v [m/s]	CoG [rad]
Epoch α	398859.55	5580658.1	2.29	4.9394563
Epoch β	398857.31	5580658.9	2.36	4.9179888
Epoch γ	398855.05	5580659.4	2.33	4.8958231
Epoch δ	398852.82	5580660.0	2.30	4.8743555
Epoch ϵ	398850.52	5580660.6	2.37	4.8532371

in this section of the river; therefore their positions and directions of two consecutive way points were taken as the values for $(\mathbf{x}_f, \mathbf{y}_f)$ and $\mathbf{CoG}_f$ for the definition of the optimal control problem. Thus, the following scenario is proposed: it is required to find the optimal

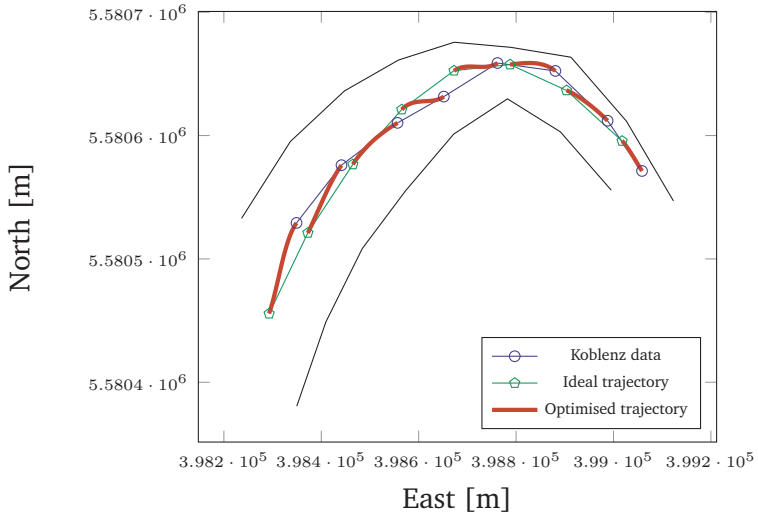


Figure 5: Way point Approaching: Koblenz PNT-Data

trajectory that will move the vessel state from $[\mathbf{E}_k, \mathbf{N}_k, \mathbf{v}_k, \mathbf{CoG}_k]'$ to $[\mathbf{E}_w, \mathbf{N}_w, \mathbf{v}_k, \mathbf{CoG}_w]'$, minimising time, acceleration and rudder movements. Here the subscripts k and w stand for Koblenz Data and the closest way point. Notice that it was assumed that the final velocity will be equal to the initial one, and that $\mathbf{CoG}_w$ is calculated as the direction of two consecutive way points.

Figure 5 illustrates the behaviour of the calculated trajectories, while Table 5 displays some of the technical characteristics of the obtained solutions.

Table 5: Parameters of the solutions using OCP

	Point							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Terminal Time [s]	53.5	34.6	31.4	30.8	38.3	29.8	36.4	35.4
Solution Time [s]	2.86	3.85	2.50	2.40	1.45	1.32	2.30	3.009
No. of Iterations	56	106	92	59	55	51	50	82

From Figure 5 and Table 5 it can be seen that the calculated trajectories follow the desired path while minimising effort and time for the navigation, with a smooth transition between the initial and final points which guaranties their feasibility when manoeuvring. Although the figure only depicts the results for the a small sample of the actual trajectory, it is worth to mention that in the set of ca. 400 points more than 95% of the solutions were satisfactory from the point of view of the convergence of the algorithm, providing optimal solutions in terms of numerical stability and computational effort. The depicted points comprise the different phases of the route, spanning a variety of scenarios that represent the difficulties of this problem and the proposed method is successful in calculating them. It is notorious that the performance of the investigated OCP method is applicable for near-real-time applications due to the low latency of its solution: 2-4 seconds per point, relatively high for real-time applications such as the PiloNav manoeuvre guidance system. Nevertheless, since the number of steps to calculate each optimal trajectory has been set arbitrarily meeting only mathematical requirements, the possibility to reduce the load of calculation and therefore to increase its performance by reducing those intermediate steps is still latent and has to be considered. Moreover, the values selected for the final state of the system should be tailored to the different stages required to navigate the channel. Thus, for instance, the determination of the final value for the $\mathbf{CoG}_w$ could be calculated using not only the direction between to consecutive points but by using a smoother curve, such as a Bezier Curve or a Spline, between three or more successive waypoints, allowing a simpler transition between the initial and final states and theoretically contributing to cut down the time of calculation.

5 Conclusion and Future Work

This work introduced a manoeuvre guidance system for inland water vessels based on highly accurate Position, Navigation and Timing (PNT) data. Where the manoeuvre guidance system consists of multiple modules, this article discusses the computation of optimal trajectories. Therefore, different evasion manoeuvres have been simulated considering static obstacles or oncoming vessels.

Two promising mathematical optimisation methods have been investigated with respect to performance and computational effort. For the presented project PiloNav the Optimal Control Problem (OCP) solver based on Sequential Quadratic Programming (SQP) has been used to be implemented into the system.

In a next step this approach has to be implemented into the DLR IKN real-time framework. In order to meet the requirements for real-time applications, the solution time has to be reduced significantly. Therefore the reduction of the number of steps in the calculated trajectory along with the smoothing of the ideal path will be investigated in more detail. Additionally, it can occur that slower or moored vessels have to be overtaken. Therefore it has to be investigated under which circumstances an overtaking of these vessels can be calculated using the presented OCP and SQP.

Moreover, more complex cost functional shall be analysed with the purpose of incorporat-

ing additional meaningful variables. River currents, wind, and the vertical component will be acknowledge with data from future measurement campaigns with the aim of coping more demanding navigation scenarios and address a broader kind of maritime applications.

Acknowledgement

This work has been supported under grant number 19G10015A (keyword: PiloNav) by the Federal Ministry of Economics and Technology (BMWi) on the basis of a decision by the German Bundestag.

References

[Alb13] T. ALBRECHT, K. LÜDDECKE, and J. ZIMMERMANN: "A Precise and Reliable Train Positioning System and its Use for Automation of Train Operation". In: *IEEE International Conference on Intelligent Rail Transportation* (2013). Bejing, China.

[Fab08a] B. C. FABIEN: "Direct optimization of dynamic systems described by differential-algebraic equations". In: *Optimal Control Applications and Methods* 29 (2008), pp. 445–466.

[Fab08b] B. C. FABIEN: "Implementation of a Robust SQP Algorithm". In: *Optimization Methods and Software* 23 (2008), pp. 827–846.

[Fab12] B. C. FABIEN: "Piecewise Polynomial Control Parameterization in the Direct Solution of Optimal Control Problems". In: *ASME J. Dynamic Systems Measurements and Control* (2012).

[Fab13] B. C. FABIEN: *OCP: An optimal control problem solver. OCP v1.0 documentation*. (Last Access April 2013). URL: `http://abs-5.me.washington.edu/ocp/`.

[Fab98] B. C. FABIEN: "Some tools for the direct solution of optimal control problems". In: *Advances in Engineering Software* 29.1 (1998), pp. 45–61.

[IMO07] *NAV 53/13 Development of an e-Navigation Strategy*. Tech. rep. International Maritime Organisation, 2007.

[IMO07a] *Resolution MSC.252(83): Adoption of the Revised Performance Standards for Integrated Navigation Systems (INS)*. Tech. rep. International Maritime Organisation, 2007.

[IMO09] *NAV 55/WP.5 Development of an e-Navigation Strategy Implementation Plan*. Tech. rep. International Maritime Organisation, 2009.

[IRI13] *Implementation of River Information Services in Europe*. `http://www.iris-europe.net/`. (Last Access 2nd June 2013). IRIS Europe Project, 2013.

[Lan13] L. LANÇA, P. ZACHHUBER, and A. BORN: "GNSS-Integrity Assessment of an Integrated PNT-Unit in a Signal Degraded Inland Water Environment". In: *PIANC - Smart Rivers 2013 Conference* (2013). Liege, Belgium.

[Lut11] A. LUTZ: "Kollisionserkennung und -vermeidung auf Binnenwasserstraßen". ger. PhD thesis. Universität Stuttgart, 2011.

[Mie03] A. MIELE and T. WANG: "Multiple-Subarc Gradient-Restoration Algorithm, Part 1: Algorithm Structure". In: *Journal of optimization theory and applications* 116.1 (2003), pp. 1–17.

[Mie05] A. MIELE and T. WANG: "Maximin Approach to the Ship Collision Avoidance Problem via Multiple-Subarc Sequential Gradient-Restoration Algorithm". In: *Journal of optimization theory and applications* 124.1 (2005), pp. 23–53.

[Mie06] A. MIELE and T. WANG: "Optimal Trajectories and Guidance Schemes for Ship Collision Avoidance". In: *Journal of optimization theory and applications* 129.1 (2006), pp. 1–21. DOI: 10.1007/s10957-006-9051-6.

[Mie74] A. MIELE, J. N. DAMOULAKIS, J. R. CLOUTIER, and J. L. TIETZE: "Sequential Gradient-Restoration Algorithm for Optimal Control Problems with Nondifferential Constraints". In: *Journal of optimization theory and applications* 13.2 (1974), pp. 218–255.

[Mie99] A. MIELE, T. WANG, C. S. CHAO, and J. B. DABNEY: "Optimal Control of a Ship for Course Change and Sidestep Maneuvers". In: *Optimal Trajectories and Guidance Schemes for Ship Collision Avoidance* 103.2 (1999), pp. 259–282.

[Rao10] A. V. RAO, D. A. BENSON, C. L. DARBY, M. A. PATTERSON, C. FRANCOLIN, I. SANDERS, and G. T. HUNTINGTON: "Algorithm 902: GPOPS, A MATLAB Software for Solving Multiple-Phase Optimal Control Problems Using the Gauss Pseudospectral Method". In: *ACM Transactions on Mathematical Software* 37.2 (2010), p. 22. DOI: 10.1145/1731022.1731032.

[Tze98] C. Y. TZENG: "Collision Avoidance by a Ship with a Moving Obstacle: Computation of Feasible Command Strategies". In: *Journal of optimization theory and applications* 97.2 (1998), pp. 281–297.

[Vie12a] I. VIERHAUS, A. BORN, and E. ENGLER: "Trajectory Optimization for Inland Water Vessels based on a next generation PNT-Unit". In: *6th ESA Workshop on Satellite Navigation Technologies* (2012), pp. 1–7.

[Vie12b] I. VIERHAUS, A. BORN, and D. MINKWITZ: "Challenges on PNT-Unit and driver assistence systems in inland water". In: *14th IAIN World Congress 2012 Seamless Navigation, IAIN 2012* (2012).

[Wac06] A. WÄCHTER and L. T. BIEGLER: "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming". In: *Journal Mathematical Programming* 106 (1 2006), pp. 25–57. ISSN: 1436-4646, 0025-5610. DOI: 10.1007/s10107-004-0559-y.

[Yav97] Y. YAVIN, C. FRANCOS, T. MILOH, and G. ZILMAN: "Collision Avoidance by a Ship with a Moving Obstacle: Computation of Feasible Command Strategies". In: *Journal of optimization theory and applications* 93.1 (1997), pp. 53–66.

Corresponding author: Dr. Alexander Born, German Aerospace Center, Institute of Communications and Navigation, Department of Nautical Systems, 17235 Neustrelitz, Germany, Tel.: +49 3981 480 219, alexander.born@dlr.de

Advantages and Limitations of a Dual Approach in Video-Based Traffic Data Acquisition

Jan Grimm
Fraunhofer Institute for Transportation and Infrastructure Systems IVI

Abstract

For traffic data acquisition and automatic incident detection, video detection has become a reasonable alternative to established detection technologies, especially due to a potential for more flexible application and lower installation costs. However, achieving a high level of quality and reliability in video-based traffic data acquisition is still a challenge in research. The objective of this paper is to present the general approach as well as advantages, practical limitations and implementation issues of the video-based algorithms developed by the Fraunhofer Institute for Transportation and Infrastructure Systems (Fraunhofer IVI). The traffic data acquisition algorithms are based on a dual approach, combining the tripwire method for vehicle counting and classification, and the tracking method for speed measurement. At first, the relevant use cases and requirements for video-based traffic data acquisition are identified. Then, the Fraunhofer video detection algorithms are described, and advantages as well as limitations found during field operational tests and in practical implementations are summarised. In addition, approaches for an online and offline quality evaluation are discussed, considering both the aspect of data quality and system performance. Finally, an outlook is given on what still needs to be done to ensure a sufficient level of quality even under adverse conditions and to convince traffic engineers and decision makers of the benefits of video detection compared to conventional traffic data acquisition techniques.

Keywords: video detection, traffic data acquisition, incident detection, image processing, quality evaluation

1 Introduction

Image processing has become an indispensable instrument in various fields of application, including automated quality control, security, and medical technology. In recent years, video detection has also gained importance for road traffic applications. Even if limited to this context, video detectors may be used for a wide range of tasks, including object detection and

identification, incident detection, and traffic data acquisition. This paper concentrates on the latter, i.e. video-based traffic data acquisition. In this context, one of the key advantages of video detectors is that installation is much more flexible, less costly and does not require pavement intrusion.

However, video-based traffic data acquisition is a difficult task. In opposition to many other video detection applications, surrounding conditions, such as lighting and weather conditions, or distance and alignment of objects to be detected, cannot be controlled. Therefore, fulfilling the requirements of real-world traffic information and traffic control applications is still a matter of research and development.

The aim of this paper is to present an efficient, robust and field-proven approach of video-based traffic data acquisition. This includes a discussion of advantages and practical limitations, a summary of related quality evaluation approaches, and a brief description of current and planned research and development activities.

2 Use Cases and Requirements of Video Detection

Traffic data is the base of various applications and processes, including real time traffic information, control and management. In addition, archived traffic data is a valuable source of information for planning, control optimisation, research, and quality assurance. The remainder of this section focuses on traffic information and traffic management, as these rely on online traffic data provision and thus have higher requirements than the remaining offline use cases.

In traffic information systems, it is usually sufficient to classify the current traffic situation in levels of service. This includes the identification of congested areas. The resulting information may be visualised on a map, or may be used to generate congestion messages to be distributed through the traffic message channel. In a traffic control application, macroscopic traffic characteristics like traffic flow or mean speed are used for automatically or semi-automatically applying certain traffic control strategies like collective re-routing, speed limits, local congestion warning, or altering the signal programmes of a traffic light intersection or corridor.

The quality of traffic information and traffic control, however, depends on the availability of traffic detectors. This is especially true if a short-term or mid-term prediction of the traffic situation shall be provided. Model-based approaches to estimate traffic flow and/or traffic distribution in a given road network may help to improve information and prediction quality. However, such models also rely on traffic detector data, and the fewer detectors are used to support the model, the more will detection errors have an impact on the resulting state estimation.

Conventional detection technologies like inductive loops are too costly for a wide-spread use throughout the road network. Their installation is only justified at certain strategic locations or where needed for traffic control applications. Video detectors, however, are available at much lower cost and thus suitable to fill detection gaps at a much better cost-benefit ratio. In addition, as the installation of video detectors does not require pavement

intrusion, video detectors can be easily adapted to changes in infrastructure. For instance, they can be used as temporary replacement at construction sites, where lane shifts may put conventional detectors out of service. At such bottlenecks, keeping traffic detection functional is of particular importance. Furthermore, in opposition to other traffic detector technologies, video detectors may be configured to provide live images of the traffic scene to a traffic control centre. This enables operators, for example, to identify the cause of an incident, or to verify the outputs of the video detection algorithms.

Both traffic information and traffic control applications have the following basic requirements concerning traffic data acquisition:

- Traffic data must be available and ready to be processed in real time.
- Traffic data must be correct, i.e. must match the actual traffic situation.
- Traffic data acquisition has to be robust, i.e. must work under all traffic and environmental conditions.

In the context of video detection, the latter point is of particular importance. A major disadvantage of video detection compared to other traffic detector technologies is its sensitivity to adverse weather conditions such as precipitation and fog, poor lighting at night, or blinding at dusk or dawn. In addition, in case of congestion or dense traffic, objects may mask one another, or multiple separate objects may appear to be merged into one. However, for the use in real-world applications, it is not sufficient for video detection algorithms to work under ideal conditions only.

From an operational point of view, a traffic detector should require as little maintenance effort as possible. The system should allow remote access, and should be able to restart itself in case of failure.

If live images of the video detectors are transmitted to a traffic control centre or even published as webcam images on a traffic information platform, privacy issues have to be taken into consideration, as well. Measures should be taken to avoid recognisability of individuals, license plates or similar details.

3 The Fraunhofer Video Detector

Fraunhofer IVI has more than ten years of experience in developing and operating video detection systems for road traffic applications. The Fraunhofer video detector can be used to obtain local macroscopic traffic characteristics such as traffic flow, mean speeds, occupancy, and queue lengths. It is suitable for both urban and motorway applications. The following description is based on [Blo05], [Rys13a] and [Rys13b].

3.1 System Architecture

The Fraunhofer video detection system can be divided into three levels: (1) detector level, (2) server level and (3) client level.

The *detector level* comprises local processing units and one or more digital or analogue

cameras per unit. For each camera, multiple detection zones can be configured, each of which usually represents a single lane of a roadway segment. A detection zone can be more than 100 metres long and may be configured to follow curved trajectories. The local processing units run the algorithms to generate real-time traffic data, which are described in subsection 3.2. Traffic data and sample images may be submitted to the server, and access from the server to the local processing unit is available for remote maintenance purposes. The bidirectional connection to the server can be realised by means of mobile or network communications; typically, UMTS is used.

The *server level* contains a central data repository for traffic data and sample images, and enables communication to the detector, to the client, and between these two levels. Configuration parameters are stored at server level, as well.

The *client level* comprises a viewer application and a configuration tool. The viewer application is a flexible means to visualise traffic data of the server database. The data can be presented either as a time series or in a fundamental diagram. The user can choose between various aggregation levels, and integrate multiple data sources into one diagram. The configuration tool is used to configure detection zones for a camera or to set detector parameters.

3.2 Algorithm Description

The basic principle of the Fraunhofer algorithm is to detect and track moving objects within pre-defined detection zones and to count them at certain tripwire locations. By combining the tracking and tripwire methods, the advantages of each individual method can be exploited, while some of their disadvantages are compensated. This dual approach, including image pre-processing steps, is described below.

For each image acquired out of a video sequence, the Fraunhofer algorithm conducts a series of processing steps for object detection, which are illustrated in figure 1.

In steps 1 to 4, the image entropy, i.e. the amount of information contained in an image, is reduced until a clear, binary differentiation between objects and background is achieved. Through the *extraction* of the detection zone subareas from the overall image (step 1), all pixels outside the detection zones are excluded from further processing. The resulting images are still distorted, i.e. remote objects appear to be smaller than those closer to the camera, even if they are equal in size. Thus, *normalisation* (step 2) is used to transform the scene from a camera-based to a world-based reference system. The next step, *segmentation* (step 3), is needed to separate foreground (i.e. objects) from background. First, the background is adaptively estimated based on recent previous images. Then, the foreground image is generated by subtracting the background image from the normalised original image. Through *binarisation* (step 4), the resulting image is converted from grey scale values to binary values, i.e. black or white, by means of a threshold value. The *dilatation* operator (step 5) is then applied to fill gaps within an object or to merge objects that are located very close to each other. As the dilatation operator has the side effect of increasing an object‘s size, the *erosion* operator (step 6) is used to restore its initial size.

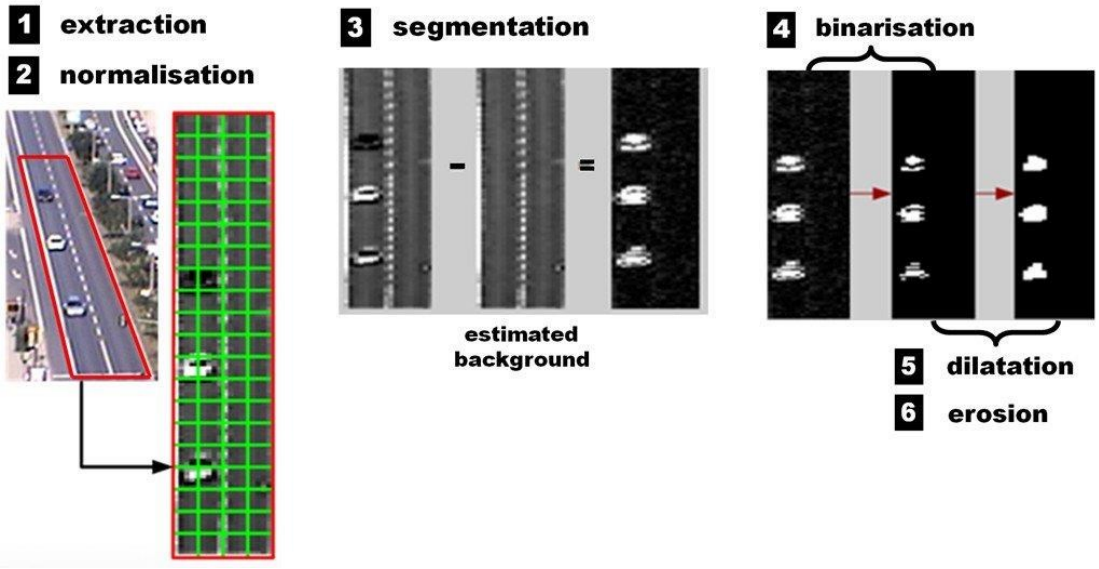


Figure 1: Processing steps for object detection within the Fraunhofer video detector.

For each object detected in a video image, certain features are extracted, particularly the object width and length. Too small objects are considered noise and excluded from further processing. The remaining objects are matched to those of the previous image by comparing their characteristics and assuming a maximum speed and a fixed driving direction as plausibility criteria. Through the spatial shift of an object's area centre and the known frame rate, the speed at which the object moved along the detection area can be calculated. If such a movement passes across the tripwire line, the vehicle is counted. This process is illustrated in figure 2.

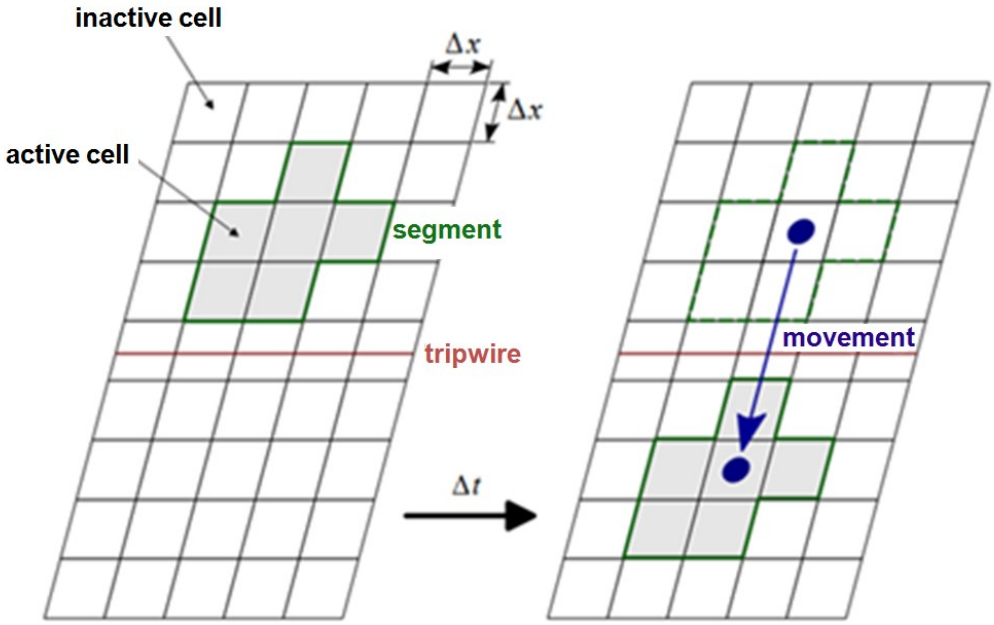


Figure 2: Speed measurement and vehicle counting with the Fraunhofer video detector.

Single speeds and vehicle counts are aggregated to one-minute intervals

- by accumulating the number of vehicles that passed the tripwire while resetting the counter to zero at the beginning of each interval and
- by calculating the average speed as arithmetic mean of the speeds represented by each movement vector within the detection area and the time interval.

In addition to vehicle counting and speed measurement, the occupancy of the detection area is estimated as the percentage of active cells out of all cells of a detection zone. Active cells, in this context, are cells which are at least partly occupied by an object.

To support traffic control or optimisation at signalled intersections, the queue length, i.e. the distance from the rear of the last stopped vehicle to the stop line, may also be calculated. In this case, the stop line position within the detection area has to be configured. The queue length value should only be used if the entire queue remains within the detection area in most cases.

3.3 Field Implementations

The Fraunhofer video detection system has been implemented on urban streets (Dresden and Nuremberg) as well as on motorways (North Rhine-Westphalia).

The Dresden system was established within the intermobil project (1999-2005). It became part of the VAMOS traffic management system and currently comprises 17 camera locations.

The Nuremberg system emerged from the ORINOKO project (2004-2008) and comprises five camera locations. The video detectors were used to alert the operators of the local traffic control centre in case of congestions by means of the eWatcher tool. In addition, traffic data of the video detectors was used to optimise control programmes for certain signalled intersections [Nur10].

By order of the North Rhine-Westphalia road administration authority (Landesbetrieb Straßenbau Nordrhein-Westfalen), around 90 cameras have been installed at strategic locations throughout the motorway network of North Rhine-Westphalia. Installation and configuration of the camera network was completed by July 2013. The system is used both for webcams and as traffic data source for a traffic information system.

At urban locations, the cameras and local processing units are typically attached to traffic signal standards or lamp poles (figure 3a), or mounted on top of large buildings like office towers (figure 3b). On motorways, most cameras are installed at walkable gantries and positioned above the median for optimal view in both directions (see figure 3c).

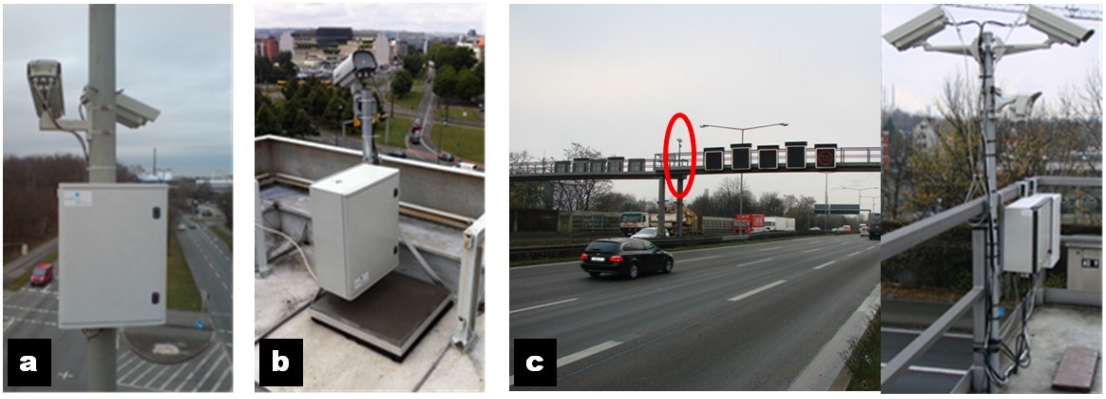


Figure 3: Examples of video detector locations.

3.4 Advantages and Practical Limitations

Through field implementations, the capabilities and advantages of the Fraunhofer video detector were demonstrated. In addition to the general advantages of video detectors over conventional detection technology described in section 2, the Fraunhofer video detector satisfies the requirements of real-world applications through:

- compatibility with various camera types, including low-cost and low-resolution cameras (which may be needed for privacy issues),
- limited maintenance effort through remote access and restart on failure,
- sufficient detection quality even in dense traffic and for most weather and lighting conditions as well as
- integrated traffic state analysis and congestion detection.

However, some practical limitations of the Fraunhofer algorithms remain, including:

- quick merge of stopped vehicles into the background due to adaptive background estimation,
- reduced quality of speed measurement at night,
- limited capabilities for vehicle classification; no vehicle classification at night,
- no wrong way driver detection and
- incident detection limited to congestion detection.

These issues are the matter of current research and further development. Section 5 provides an overview on current activities.

4 Quality Evaluation Approaches

In the context of video-based traffic data acquisition, quality evaluation is necessary for the following purposes:

1. To identify and flag data for which a reduced level of quality can be expected, e.g. due to adverse weather conditions,
2. To identify and eliminate systematic errors, e.g. attributed to sub-optimal detector configuration,
3. To identify accumulations of device, communication or software failures which need attention of maintenance personnel.

To mark data for which a low level of quality is assumed, a quality index is attributed to each one-minute interval data record. The quality index is based on an estimation of the noise component of the image signal. To improve the significance of the quality index, plans for further development include environmental conditions to be considered, as well.

To identify systematic errors, a combination of random plausibility checks and reference data comparison is a well-tried approach. As some systematic errors occur only under certain conditions, it is useful to conduct such evaluations individually for different situation classes. Situation classes, in this context, may be characterised by time of the day, weather conditions, and traffic situation. The reference data may be obtained through manual evaluation of sample video sequences, or – if available – from conventional traffic detectors (e.g. inductive loops) located within or close to the detection area of the video detector under test.

However, in more complex video detection systems, this approach may prove to be too time-consuming. Thus, it is only applicable for algorithm or configuration tests, or for tracking down the cause of issues already identified. Likewise, monitoring and tracking hardware and software failures manually may turn out to be impractical. For this reason, within the Traffic IQ project (2010-2013), Fraunhofer IVI developed and implemented an automated tool to detect and visualise errors and failures in complex video detection systems. This tool is even capable of automatically handling certain issues (e.g. rebooting a local processing unit on hardware failure). For quality management purposes, the system performance may be summarised in availability, timeliness and correctness parameters, which may be aggregated by time, region and infrastructure hierarchy level [Ruh13].

5 Outlook: Future Research Activities

As discussed in section 3.4, the Fraunhofer video detector is ready to be used in real-world applications, but there is still a potential for research and further development. This section discusses three examples of current research and development activities in detail.

Currently, a high priority is assigned to the improvement of vehicle classification. The existing algorithm is able to distinguish passenger cars from large vehicles (lorries, buses) using the vehicle length and width. However, classification quality proved to be insufficient under certain circumstances, particularly at night. In addition, for some applications, a classification of more than these two vehicle types is desired. As a first step, a multi-criteria maximum likelihood approach will be considered. Another approach for this task might be using pattern recognition and artificial intelligence. However, due to the variety of camera scenes, objects to be detected and surrounding conditions, such an algorithm will have to be

optimised for each camera individually, which will result in a high effort required for camera implementation. Thus, the former approach is preferred.

Another field of activity is the background estimation algorithm. To prevent stopped vehicles from merging with the background, the potential for further development of this algorithm is evaluated. Hereby, stopped vehicles can be kept detectable for a longer time. However, the cause of the problem will remain. Thus, the vehicle tracking algorithm is planned to be expanded by integrating a simple traffic flow model and a Kalman Filter as a state predictor.

Plans for future research activities also include the integration of automatic incident detection into the Fraunhofer video detector. In this context, automatic wrong way driver detection is of particular interest. Other detectable incidents include single stopped or slow-moving vehicles, or accumulations of abrupt manoeuvres such as hard braking or swerving, just to name a few examples.

In addition to these activities, Fraunhofer IVI continuously works on further improving the existing algorithms and quality evaluation tools in close co-operation with field experts using the Fraunhofer video detector.

References

[Blo05] S. Blobel: “Untersuchungen zur makroskopischen Verkehrszustandsanalyse”. Diploma thesis. Dresden, Germany: University of Applied Sciences, 2005.

[Nur10] City of Nuremberg (Ed.): “ORINOKO – Abschlussbericht”. Final project report. Nuremberg, Germany, May 2010.

[Ruh13] S. v. d. Ruhren et al: “Traffic IQ – Pilotprojekt Informationsqualität im Verkehrswesen”. Final project report. Aachen, Germany, 2013.

[Rys13a] T. Ryssel: “Fraunhofer IVI Videodetektion – Handbuch zum Konfigurationstool”. Reference manual. Dresden, Germany, 2013.

[Rys13b] T. Ryssel: “Fraunhofer IVI Videodetektion – Handbuch zum Statistiktool”. Reference manual. Dresden, Germany, 2013.

Corresponding author: Jan Grimm, Fraunhofer Institute for Transportation and Infrastructure Systems IVI, 01069 Dresden, Germany, phone: +49 351 4640 620, e-mail: jan.grimm@ivi.fraunhofer.de

A Comparative Study of Shadow Models for Video-Based Traffic-State Analysis

Klaus-Peter Döge
Technische Universität Dresden

Abstract

The recognition of shadows is an important object of research in the field of image processing. First of all in video-based traffic-state analysis, shadows cause significant disturbances. Today, shadow modelling and detection has been rather extensively examined, but there is a lack of systematisation and evaluation of basic concepts. Therefore, this paper contains a short introduction of the basic concepts and the use of these concepts in reliable methods. To get a comparative study, every method is applied to the same example images. The study points out the need for further investigation of basic technologies and applications.

Keywords: shadow modelling, shadow identification, image processing, photometric invariants, colour models, video-based traffic-state analysis

1 Introduction

From an automatic image processing standpoint, shadows are changing the image content essentially and influencing the detection results negatively. In this paper, we focus on moving traffic, neglecting stationary traffic like parking and accidents. The main problems caused by shadows are: 1. Confusion of shadows and vehicles in recognition and tracking tasks. 2. The merging of vehicles and shadows, making classification more difficult or even impossible. 3. A general darkening of scenes, hindering the automatic image processing by increasing the contrast. The shadow itself occurs in two types. The first is the so-called cast-shadow, which is the projection of an object on to a surface because of the occlusion of a light source. The second is the self-shadow, a type of shadow that appears on the side of an object turned away from the light. Figure 1 shows some examples of these two types of shadows. In the case of strong self-shadows, the darkened part of the vehicle merges with the cast-shadow. So figure 1 also illustrates the main challenge: The separation of cast-shadow and self-shadow. The following requirements for *every* shadow model or algorithm can thus be defined [Doe13]: I. Recognition of cast-shadows, II. Ignoring self-shadows to retain most of the vehicle. III. Recognition of “no-shadows” to avoid false detection of dark parts.

These requirements are used in this study to evaluate and compare the different methods. The basic method for fulfilling these ambitious requirements is the so-called “photometric

invariant": The road can be covered by a vehicle or a shadow. An invariant is a measurement based on the non-covered part of the road. In the case of the shadow-covered road the measure changes its value only slightly, while in the case of vehicle-covered road the change is significant. Typical invariants are the "Normalised RGB", which relates to the corresponding colour model [Bun05][Fin02], and the "Reflection Ratio" [Nay93]. Both are described in this paper along with some applications. The results are summarised in table 1 and table 2 to give an idea how each method fulfils the requirements defined above. This is only an overview of extensive material; for a more detailed description in English, please refer to the indicated sources or to [Doe13] in German.

2 Basic concepts

For the comparative study we have used three example images (figure 1). Instead of the image itself, a difference image is primarily used to extract dynamic image content like vehicles and vehicle shadows. Basic methods for distinguishing between the last two entities supply the "photometric invariants". The result is a "shadow mask" with black pixels for the shadow and white pixels for the rest of the image content.

Figure 1: The example images of the comparative study. Left: a coloured (red) vehicle with a strong self-shadow and cast-shadow. Middle: a white vehicle with a cast-shadow and a light self-shadow. Right: a white vehicle with a light cast-shadow and a light self-shadow.

2.1 Colour modelling

For the modelling of captured RGB components the following idea has been established in a number of variants [Bun05][Fin02][Fun95][Gev99][Sal95], while the following parameters – depending on the wavelength λ– are considered: sensitivity of the camera sensor $\sigma_x(\lambda)$, spectral reflectance of the surface $\varrho(\lambda)$ and the spectrum of the light source $E(\lambda)$.

The reaction of the RGB sensor is calculated along the whole range of wavelengths Λ:

$$b_x = \int_\Lambda \sigma_x(\lambda)\varrho(\lambda)E(\lambda)\,d\lambda,\ x \in \{r,g,b\}. \tag{1}$$

The sensors possess different sensitivities in their own spectral ranges, so one can expect that each sensor is responsible for exactly one wavelength. With this in mind, the equation (1)can be simplified, resulting in the basic equation for shadow modelling with the help of the so-called "invariants":

$$b_{x,x\in\{r,g,b\}}(\lambda) = \sigma_x(\lambda_x)\varrho(\lambda_x)E(\lambda_x). \tag{2}$$

1.1 The "Normalised RGB" invariant

The spectral radiance depending on the wavelength λ is described by Planck's radiation law with the parameters of Planck's constant h, the velocity of light c, the solid angle Ω_0, Bolzmann's constant k and the absolute temperature T:

$$E(\lambda) = C_1\lambda^{-5}\left(e^{\frac{C_2}{\lambda}} - 1\right)^{-1}, C_1 = \frac{2hc^2}{\Omega_0}, C_2 = \frac{hc}{\Omega_0 kT}. \quad (3)$$

Additionally, the parameter $\alpha = (0..1)$ models the darkening of every RGB component by a shadow. Inserting equation (3) into equation (2) and adding the parameter α gives

$$b_{x,x\in\{r,g,b\}}(\lambda) = \sigma_x(\lambda_x)\varrho(\lambda_x)C_1\lambda^{-5}\left(e^{\frac{C_2}{\lambda}} - 1\right)^{-1}\alpha. \quad (4)$$

Using equation (4) in the normalised RGB colour model [Bun05], the result is independent from the darkening, but still depends on the camera and background parameters. Therefore, "Normalised RGB" is a useful invariant for shadow detection [Fin02]:

$$b_{x_{norm}} = \frac{\sigma_x(\lambda_x)\varrho(\lambda_x)C_1\lambda^{-5}\left(e^{\frac{C_2}{\lambda}} - 1\right)^{-1}}{\sum_{i\in\{r,g,b\}}\sigma_i(\lambda_i)\varrho(\lambda_i)C_1{\lambda_i}^{-5}\left(e^{\frac{C_2}{\lambda_i}} - 1\right)^{-1}}, x \in \{r,g,b\}. \quad (5)$$

2.2 The "Reflection Ratio" invariant

The "Reflection Ratio" invariant is widely used and is described for instance in [Nay93]. It is based on the relation of the spectral reflectance $\varrho(\lambda)$ of two points on a surface. Applying equation (2) for the two points $b^{p1}_{x,x\in\{r,g,b\}}$ and $b^{p2}_{x,x\in\{r,g,b\}}$ leads to:

$$b^{p1}_{x,x\in\{r,g,b\}} = \sigma_x(\lambda_x)\varrho^{p1}(\lambda_x)E^{p1}(\lambda_x), \qquad b^{p2}_{x,x\in\{r,g,b\}} = \sigma_x(\lambda_x)\varrho^{p2}(\lambda_x)E^{p2}(\lambda_x). \quad (6)$$

The assumptions that the points are near together, so that $E^{p1}(\lambda_x) \approx E^{p2}(\lambda_x)$, and the sensor sensitivities $\sigma_x(\lambda_x)$ are almost equal, leading to the formulation of the invariant in equation (7), while the variety ϱ^* avoids division by zero.

$$\varrho = \frac{b^{p1}}{b^{p2}} = \frac{\varrho^{p1}}{\varrho^{p2}}, resp. \quad \varrho^* = \frac{b^{p1} - b^{p2}}{b^{p1} + b^{p2}} \quad (7)$$

3 Application of the basic concepts

3.1 Quotient image with smoothing

The idea of the quotient image is based on the assumption that shadow parts of an image are always darker than the rest of the image. In [Als04][Bev03] one can find the justification that the relation between shadow and background intensity values is almost constant. Finally, it is an application of the invariant "Reflection Ratio". The quotient image $\boldsymbol{B}_{quot}$ can be calculated from a reference image $\boldsymbol{B}_R$ and the image to be analysed $\boldsymbol{B}_A$ where the operator $\oslash$ denotes

the piecewise division of the intensity values. The images $\boldsymbol{B}_R, \boldsymbol{B}_A$ and $\boldsymbol{B}_{quot}$ are smoothed with a Gauss-Filter of third order.

$$\boldsymbol{B}_{quot} = \boldsymbol{B}_R \oslash \boldsymbol{B}_A \tag{8}$$

Because the intensity values in $\boldsymbol{B}_{quot}$ do not lie between 0 and 255 they are scaled

$$\boldsymbol{B}_{quot,255} = \boldsymbol{B}_{quot} \frac{255}{\max(\boldsymbol{B}_{quot})}, \tag{9}$$

and a shadow mask $\boldsymbol{B}_S$ as shown in figure 2, depending on a threshold T can be calculated.

$$\boldsymbol{B}_S = \begin{cases} 0\ (black), if\ \boldsymbol{B}_{quot,255} \leq T, \\ 1(white)\ else. \end{cases} \tag{10}$$

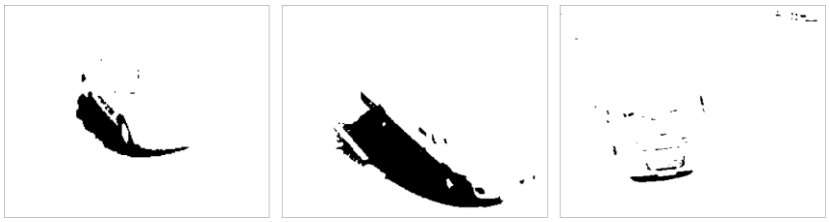


Figure 2: The shadow masks $\boldsymbol{B}_S$ for the "quotient image".

3.2 Mean and median of the image parts

Another way to find shadows automatically is the use of image parts [Sca90]. In our example, an image part contains 5x5 pixels. While μ_i is the mean of the intensity values of the *i*-th part, the median *M* is calculated from all values μ_i. With these two values a shadow mask can be created according to equation (11). The idea behind this formula is that "The mean of the shadows' intensity values is less than the median of all means". For the present example, this idea is reviewed in [Doe13]. The use of half of the median leads to an increase of robustness against noise.

$$\boldsymbol{B}_S = \begin{cases} 0\ (black), if\ \mu_i \leq M/2, \\ 1\ (white)\ else. \end{cases} \tag{11}$$

Figure 3: Shadow masks $\boldsymbol{B}_S$ for the "mean and median of the image parts".

3.3 "Reflection ratio" invariant for RGB images

The "Reflection Ratio" invariant [Bun05][Nay93] can be used for colour models with comparable information in each channel, like the RGB model. According to [Bun05] we are

using equation (7) with a fourth neighbourhood, meaning that diagonal pixels are set into relation:

$$\varrho_x{}^*(u,v,i,j) = \frac{b_x(u,v) - b_x(u+i,v+j)}{b_x(u,v) + b_x(u+i,v+j)} \quad resp. \tag{12}$$
$$\varrho^*_{R,x}(u,v,i,j) = \frac{b_{R,x}(u,v) - b_{R,x}(u+i,v+j)}{b_{R,x}(u,v) + b_{R,x}(u+i,v+j)}, x\epsilon\{r,g,b\}; i\epsilon\{-1,1\};\ j\epsilon\{-1,1\}.$$

Referring to equation (7), the invariant is calculated for every pixel in the image to be analysed ($\varrho_x{}^*(u,v,i,j)$) and a reference image ($\varrho^*_{R,x}(u,v,i,j)$). The nature of an invariant is to be constant with and without the occurrence of shadows; therefore the shadow mask $\boldsymbol{B}_S$ can be calculated from the difference of the invariants, which is less or equal to threshold T.

$$\boldsymbol{B}_S = \begin{cases} 0\ (black), if\ \sum_{x\epsilon\{r,g,b\}} \left|\varrho^*_{R,x}(u,v,i,j) - \varrho^*_x(u,v,i,j)\right| \le T, \\ 1\ (white)\ else. \end{cases} \tag{13}$$

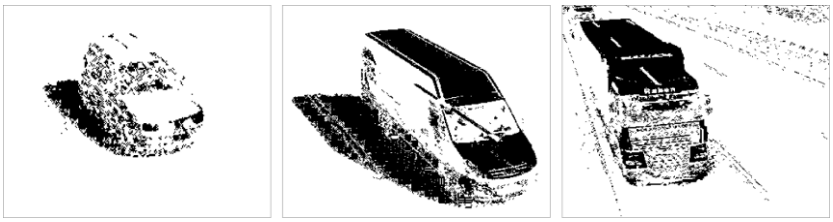


Figure 4: Shadow masks $\boldsymbol{B}_S$ for the "'Reflection Ratio' invariant for RGB images".

3.4 Statistic characteristics of the "reflection ratio" for RGB images

The idea of the "Reflection Ratio" invariant, according to [Nay93] is supplemented in various ways. For instance, [Son07] assumed constant conditions in the colour channels. So the relation $v_x(u,v)$ between RGB values on the road $b_{ro,x}(u,v)$ (without shadows) and the RGB values in shaded parts $b_{sh,x}(u,v)$ can be calculated:

$$v_x(u,v) = \frac{b_{sh,x}(u,v)}{b_{ro,x}(u,v)}, x\epsilon\{r,g,b\}. \tag{14}$$

Every $v_{x,x\epsilon\{r,g,b\}}(u,v)$ has a different mean μ_x and variance σ_x, so that the shadow mask $\boldsymbol{B}_S$ can be calculated according to equation (15). The threshold $1.5\sigma_x$ relates to the 88.6 % value of the Gaussian distribution.

$$\boldsymbol{B}_S = \begin{cases} 0\ (black), if\ \left|\frac{b_x(u,v)}{b_{ro,x}(u,v)} - \mu_x\right| \le 1.5\sigma_x, \forall x\epsilon\{r,g,b\}, \\ 1\ (white)\ else. \end{cases} \tag{15}$$

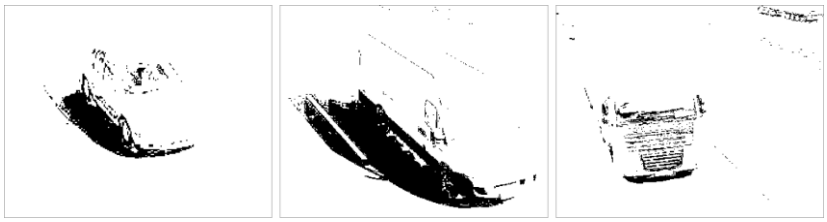


Figure 5: Shadow masks $\boldsymbol{B}_S$ for "statistic characteristics of the Reflection Ratio".

3.5 "Normalised RGB" invariant

The "Normalised RGB" invariant [Bun05][Fin02] can be applied directly to the image to be analysed ($b_{x_{norm}}$) and a reference image ($b_{R,x_{norm}}$).

$$b_{x_{norm}} = \frac{b_x}{\sum_{i\epsilon\{r,g,b\}} b_i}, resp.\ \ b_{R,x_{norm}} = \frac{b_{R,x}}{\sum_{i\epsilon\{r,g,b\}} b_i}, x\epsilon\{r,g,b\} \tag{16}$$

Due to the character of the invariant, which is independent of the occurrence of a shadow, the shadow mask $\boldsymbol{B}_S$ can be calculated by the evaluation of $\left|b_{R,x_{norm}} - b_{x_{norm}}\right|$ with the help of threshold T:

$$\boldsymbol{B}_S = \begin{cases} 0\ (black), if\ \left|b_{R,x_{norm}} - b_{x_{norm}}\right| \leq T, \forall x\epsilon\{r,g,b\}, \\ 1\ (white)\ else. \end{cases} \tag{17}$$

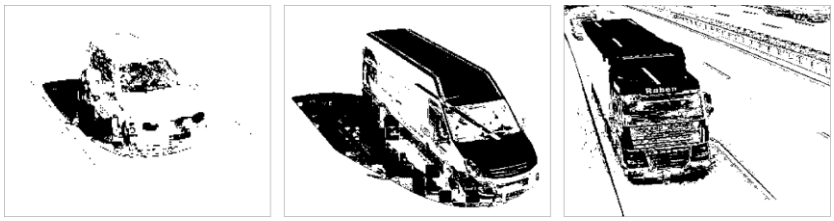


Figure 6: Shadow masks $\boldsymbol{B}_S$ for the "Normalised RGB" invariant.

3.6 Use of the HSV colour model

Independently from the variants discussed so far, the HSV colour model is also able to distinguish between vehicle, road and shadow. In [Cuc01][Cuc03] the following formula for the calculation of a shadow mask is suggested:

$$\boldsymbol{B}_S = \begin{cases} 0\ (black), if\ (\alpha \leq \frac{b_V}{b_{R,V}} \leq \beta) \wedge (\left|b_S - b_{R,S}\right| \leq \tau_S) \wedge (D_H \leq \tau_H), \\ 1\ (white)\ else. \end{cases} \tag{18}$$
$$D_H = \min(\left|b_H - b_{R,H}\right|, 360° - \left|b_H - b_{R,H}\right|)$$

It consists of three $\wedge$-connected terms. The term$(\alpha \leq \frac{b_V}{b_{R,V}} \leq \beta)$ contains the V-component of the image to be analysed (b_V) and the V-component of the reference image ($b_{R,V}$), while α is the assumed darkening, and β relates to the robustness against noise. The second and

third term ($|b_S - b_{R,S}| \leq \tau_S$) and ($D_H \leq \tau_H$) take into consideration the change of the S and H-components by the rise of a shadow.

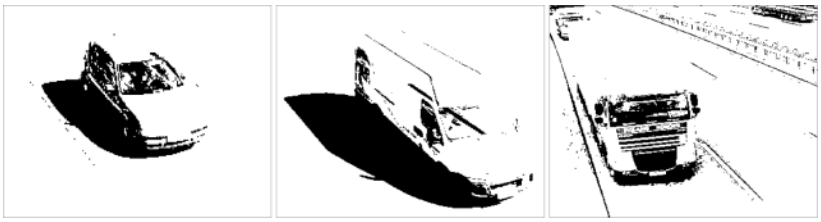


Figure 7: Shadow masks $\boldsymbol{B}_S$ for the "HSV colour model".

4 Evaluation of the basic concepts

The presented methods can be divided in two categories: The first works on the dynamic parts of the difference image (table 1) and the second on the entire image (table 2).

Table 1: Comparative evaluation of shadow models for the difference image.

Method	I. Recognition of cast-shadow	II. Ignoring self-shadow	III. No shadow in the image
"Reflection Ratio" invariant for RGB images [Bun05][Nay93]	The strong cast-shadow is completely recognised, but somewhat blurry. Very light parts of the cast-shadow are neglected.	The strong cast-shadow is recognised but blurry. The light parts of the cast-shadow are hardly recognised.	The complete top and front side of the white vehicles are incorrectly recognised as shadows. The darkening under the vehicle is partly recognised.
Statistic characteristics of the "Reflection Ratio" invariant for RGB images [Son07]	The strong cast-shadow is correctly recognised except on about 10% of the rear end of the vehicle. Very light cast-shadows are neglected.	Strong and light self-shadows are almost completely suppressed.	Darkening under the vehicle is well recognised. Some dark parts of the vehicle front are falsely detected.
"Normalised RGB" invariant [Bun05][Fin02]	The strong cast-shadow is recognised but partly incomplete within the range of the wheels and under the vehicle. The light cast-shadow is not recognised.	The self-shadow is partly recognised.	The white vehicle's tops, bonnets and car windscreens have a lot of falsely detected shadows.
Use of the HSV colour model [Cuc01][Cuc03]	Strong cast-shadows are well detected in every colour channel, particularly in the V-channel. Very light cast-shadows are not detected.	It is difficult for the HSV colour model to neglect self-shadows.	Dark areas of the image are often mistakenly recognised as shadows.

Table 2: Comparative evaluation of shadow models for the entire image

Method	I. Recognition of cast-shadow	II. Ignoring self-shadow	III. No shadow in the image
Quotient Image with smoothing [Als04][Bev03]	Most of the cast-shadow is recognized. The light part of the cast-shadow is neglected.	Strong and light parts of the self-shadow are effectively suppressed	Some dark parts of the vehicle are incorrectly identified as shadows. Darkening under the vehicle is well recognised.
Mean and median of image parts [Sca90]	Most of the strong cast-shadows are recognised. The light part of the cast-shadow is not recognised.	The strong self-shadow is misleadingly recognised. The light self-shadow is correctly recognised.	Large parts of the vehicle front are recognised incorrectly. Darkening under the vehicle is well recognised.

As a rule, strong cast-shadows are recognised well by every method. Very light self-shadows cause problems for all methods, but usually, self-shadows are partly recognised. Please note that the difference image a priori does not contain a lot of information in the area of self-shadows for the truck and the delivery van.

Compared to grey values, colour models use extended information. The results of the experiments have shown that the advantage of colour models is only applicable for brightly coloured vehicles. In any other cases, the difference between the vehicle and the road is difficult for colour models to detect.

5 Conclusion

The present paper investigates, systematises and evaluates basic technologies for shadow recognition in traffic-state images. As part of a growing number of methods, the so-called “photometric invariants” supplies the basis for solving the difficult problems of distinguishing vehicles, cast-shadows and self-shadows.

The approaches can be considered under the aspect of colour and grey value analysis. While the colour information extends the mathematical possibilities, one should remember that at night, normally greyscale images are used.

The recognition of cast-shadows can be done successfully using most of the introduced approaches. On the other hand, the recognition of self-shadows as part of the vehicle is much more complicated. Further work on this topic is necessary, making it an interesting object of research.

References

[Als04] F. E. ALSAQRE and Y. BAOZONG: "Moving shadows detection in video sequences". In: *Proceedings of the ICSP*. 2004, pp. 1306–1309.

[Bev03] A. BEVILACQUA: "Effective shadow detection in traffic monitoring applications". In: *J. Winter School of Computer Graphics* 11.1 (2003), pp. 57–64.

[Bun05] F. BUNYAK, I. ERSOY, and S. R. SUBRAMANYA: "Shadow detection by combined photometric invariants for improved foreground segmentation". In: *Proceedings of the 7th IEEE Workshop on Application of Computer Vision.* Vol. 1, 2005, pp. 510–515.

[Cuc01] R. CUCCHIARA, M. PICCARDI, and A. PRATI: "Improving shadow suppression in moving object detection with HSV color information". In: *Proceedings of the IEEE conference for intelligent transportation systems*. 2001, pp. 334–339.

[Cuc03] R. CUCCHIARA, M. PICCARDI, and A. PRATI: "Detecting moving objects, ghosts, and Shadows in video streams". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25.10 (2003), pp. 1337–1342.

[Doe13] K.-P. DÖGE: *Videodetektion im Straßenverkehr: Signalmodelle und Analyseverfahren.* Munich, Germany: Oldenbourg Wissenschaftsverlag, 2013.

[Fin02] G. FINLAYSON, S. HORDLEY, and M. DREW: "Removing shadows from images". In: *Proc. of the 7th European Conference on Computer Vision*. 2002, pp. 823–836.

[Fun95] G. FUNKA-LEA and R. BAJCSY: "Combining color and geometry for the active, visual recognition of shadows". In: *Proceedings on the 5th International Conference on Computer Vision*. 1995, pp. 203–209.

[Gev99] T. GEVERS and A. W. M. SMEULDERS: "Color-based object recognition". In: *Pattern Recognition* 32 (1999), pp. 453–464.

[Nay93] F. S. K. NAYAR and R. M. BOLLE: "Computing reflectance ratios from an image". In: *Pattern Recognition* 26.10 (1993), pp. 1529–1542.

[Sal95] E. SALVADOR, A. CAVALLARO, and T. EBRAHIMI: "Cast shadow segmentation using invariant color features". In: *Computer Vision and Image understanding* 95 (1995), pp. 238-259.

[Sca90] J. M. SCANLAN, D. M. CHABRIES, and R. W. CHRISTIANSEN: "A shadow detection and removal algorithm for 2-D images". In: *Proceedings of the International Conference for Acoustics, Speech and Signal Processing*. Vol. 4. 1990, pp. 2057–2060.

[Son07] K.-T. SONG and J.-C. TAI: "Image-based traffic monitoring with shadow suppression". In: *Proceedings of the IEEE* 95.2 (2007), pp. 413–426.

Corresponding author: Klaus-Peter Döge, Technische Universität Dresden, "Friedrich List" Faculty of Transport and Traffic Sciences, Institute of Traffic Telematics, 01062 Dresden, Germany, phone: +49 351 463 36779, e-mail: klauspeter.doege@tu-dresden.de

Camera-Assisted Passenger Localization in Public Transport Vehicles

Uwe Gosda, Richard Weber, Oliver Michler
Technische Universität Dresden

Abstract

Localization of people or objects is becoming an increasingly important task in a wide area of ITS-related applications. In this paper we address the problem of localizing people in public transport vehicles, e.g. as part of an automated ticketing system. In our application positioning information is primarily acquired by means of radio detection and ranging in a wireless sensor network (WSN). We show how to further improve the accuracy of the positioning estimate by data fusion with the vehicle's surveillance camera system. The possible role of cameras in the localization process is examined through the example of a single person that is observed from distinct viewpoints. We suggest a data fusion method that is based on combining 2D-Gaussians in the WSN's coordinate system and show how incorporating positioning information from cameras can improve the overall localization accuracy.

Keywords: Localization, Data Fusion, People Detection, Wireless Sensor Network

1 Introduction

Localization methods play an important role in many of today's ITS (Intelligent Transportation Systems) applications. In the context of this paper we concentrate on improving the localization accuracy of passengers in public transport vehicles. Our long-term goal is a fully automated ticketing system that does not require user activities such as check-in or check-out at terminals. A possible solution to this problem is the continuous localization of a radio tag on each passenger's electronic ticket in the vicinity of a public transport vehicle. In this approach the tag could be assigned to a vehicle based on its position. Our primary source of data for localization in this context is a Wireless Sensor Network (WSN) that is capable of measuring the distances between mobile radio tags and a set of fixed anchor nodes inside a vehicle. Based on the distance measurements the position of the mobile node can then be estimated, e.g. using multilateration methods [Web11]. In this application the requirements regarding positioning accuracy are quite challenging. In order to avoid false registrations it is

crucial to improve the WSN's localization accuracy. One way to achieve this goal is the incorporation of additional sensors in the positioning process using data fusion methods. While there are many possible sources of additional data (e.g. accelerometers), we concentrate on on-board cameras here. Within the scope of this paper we assume that a valid position estimation of a mobile tag is available from a radio based method (WSN localization in our case). Since the continuous tracking of multiple mobile tags in a WSN can be very time consuming we also avoid state estimators like Kalman Filters in our approach. Instead, we suggest a method based on the combination of two dimensional Gaussian distributions in order to merge all the information available at a specific point in time. Furthermore, we restrict the problem to estimating the position of a single person carrying a mobile node in his hand while being observed by two cameras viewing the scene from opposing points in the vehicle. Thus, the problem of assigning multiple mobile nodes to positions in the video frames is circumvented in this work.

This paper is structured as follows: In the following section we introduce the setup of sensors we used and as well give a brief overview of the geometry between the camera and world coordinate frames. In Section 3 we describe our data fusion approach. We outline a method for combining multiple 2D Gaussian distributions and describe the error models for different sensor measurements. Section 4 contains results of our approach in a sample scenario, followed by conclusions and future work in the last section.

2 Sensor Setup and Perspective Geometry

2.1 Sensor Setup for Localization

In this work we use a WSN as primary data source for localization. It typically consists of several spatially distributed, autonomous sensor nodes, that communicate with each other by multi-hop networking. For localization purposes the WSN can provide distance measurements between the network's individual nodes. This process is called ranging. Distances are measured between pairs of nodes by time-of-arrival, time-of-flight or phase-of-arrival ranging techniques [Lan06; Ben07]. Based on a set of distances a position can be computed using multilateration. Although this kind of localization allows for the estimation of 3D coordinates, we restrict the WSN-based localization to two dimensions by mapping all 3D positions to a common height on a plane parallel to the vehicle floor. Throughout this work we will refer to this as the WSN coordinate frame. Fig. 1c shows a map of the vehicle in this plane, including the sensor node positions, camera positions and some vehicle interior. We rely on two cameras in typical locations inside the vehicle for improving the WSN position estimation. In Fig. 1c the coordinates of the camera positions are denoted with C_1 and C_2, respectively. They observe the same scene from different viewpoints, i.e any object in the field of view of one camera is visible in the other camera's image provided it is not occluded. Fig. 1a and 1b show sample images of the cameras corresponding to the same point in time.

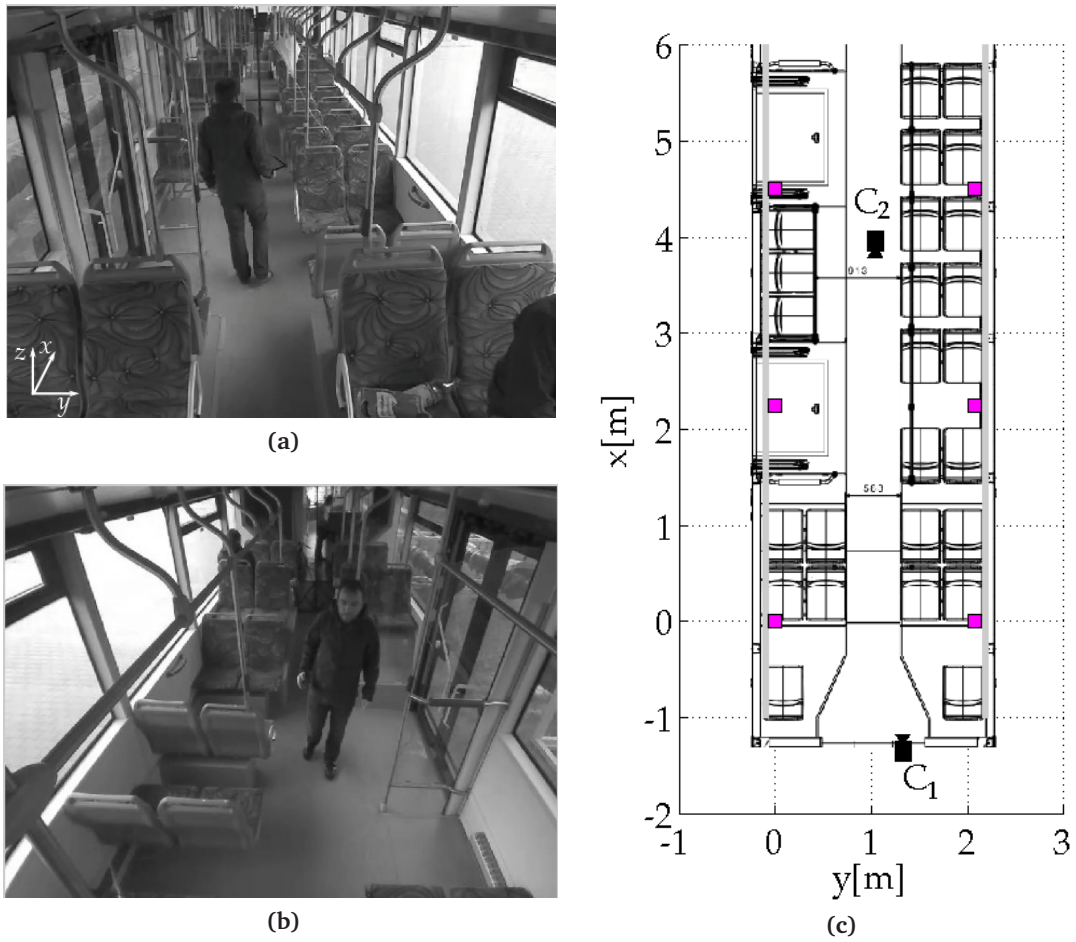


Figure 1: Sample camera images and WSN coordinate frame: (a) sample image taken from camera C_1, (b) sample image of camera C_2, (c) map of the vehicle including camera positions (C_1 and C_2) and WSN anchor nodes (squares) in the WSN coordinate frame.

2.2 Geometric Relation between Coordinate Systems

A point in the 3D world coordinate system can be mapped to a point in the camera coordinate frame using perspective transformation:

$$p_C = P p_W = KR[I \mid -C]p_W, \tag{1}$$

where points p_W are denoted as homogeneous world coordinates $(x_W, y_W, z_W, 1)^T$. The resulting point in homogeneous camera coordinates can be written as $p_C = w(u, v, 1)^T$, accordingly. It relates to the point $(u, v)^T$ in image coordinates, thus w is referred to as a scaling factor. The 3×4 projection matrix P can be decomposed into the 3×3 matrix of internal camera parameters K and the parameters R and C that relate the camera orientation and position w.r.t. the world coordinate frame. Here R is a 3×3 rotation matrix and C is the camera center in 3D world coordinates $(x_C, y_C, z_C)^T$, i.e. the displacement of the world coordinate frame's origin w.r.t. the camera position.

The operation that reverses the projection of Eq. (1) is called back-projection. A point $(u, v)^T$ in a camera image relates to a ray through the camera center and the corresponding point in world coordinates. This relation can be written as [Har04]:

$$p_W(\lambda) = \lambda P^+ p_C + C, \tag{2}$$

where P^+ is the pseudo-inverse of the projection matrix P. Thus, a point in world coordinates can be mapped to image coordinates and back using Equations (1) and (2). The position λ on the ray is then given by the scaling factor w of the forward projection, but remains unknown in the general case.

3 Data Fusion Approach

3.1 Combining 2D-Gaussians

In this paper we consider localization in the Cartesian plane given in Fig. 1c. Furthermore, we model the positioning error of the different sensor types as two dimensional Gaussian distributions in this plane. Fig. 2a shows an example of a 2D Gaussian. Its mean is the center of an ellipse whose semi-major and semi-minor axes are the standard deviations σ_{maj} and σ_{min} of a 2D Gaussian distribution. The angle θ of the semi-major axis defines the orientation of this distribution w.r.t. the WSN coordinate frame. Stroupe et.al. introduced a method for

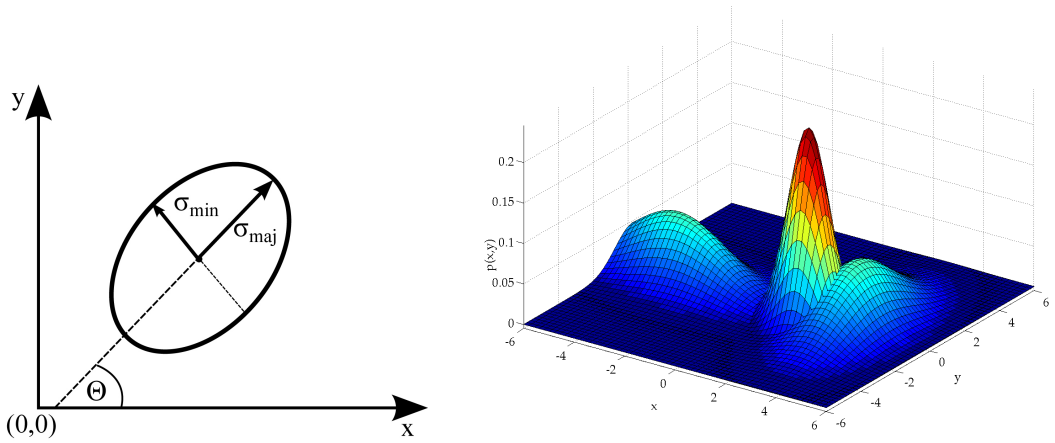


Figure 2: Merging of 2D Gaussians:(a) Parameters of a 2D Gaussian distribution, (b) plot of two 2D Gaussians (left,right) and the resulting combined Gaussian (middle)

merging multiple two dimensional Gaussian distributions using simple matrix operations in [Str00]. Their approach requires the estimation of a covariance matrix Q_L of an observation in the form of:

$$Q_L = \begin{bmatrix} \sigma_{maj}^2 & 0 \\ 0 & \sigma_{min}^2 \end{bmatrix}. \tag{3}$$

Since the orientation of Q_L with respect to the WSN coordinate system is given by the angle θ the covariance matrix can be transformed to this coordinate frame using

$$Q = R_{2D}^T Q_L R_{2D}, \tag{4}$$

where the planar rotation matrix is given by:

$$R_{2D} = \begin{bmatrix} \cos(-\theta) & -\sin(-\theta) \\ \sin(-\theta) & \cos(-\theta) \end{bmatrix}. \tag{5}$$

The combination Q' of two covariance matrices Q_1 and Q_2 can then be computed as follows:

$$Q' = Q_1 - Q_1(Q_1 + Q_2)^{-1} Q_1. \tag{6}$$

The corresponding combined mean X' is the sum of the mean X_1 and the weighted residual of the individual means X_1 and X_2 :

$$X' = X_1 + Q_1(Q_1 + Q_2)^{-1}(X_2 - X_1). \tag{7}$$

There are several ways to compute the angle θ' of the principle axis from the combined covariance matrix. An easy approach is to apply singular value decomposition, yielding σ_{maj} and σ_{min} of the merged distribution as the eigenvalues of Q'. The direction of the semi-major axis can then be computed from the corresponding eigenvectors. Fig. 2b shows the result of merging two 2D Gaussian distributions using the approach described above. The variance of the distribution function is smaller than the variances of the individual distributions as a result of their combination.

3.2 Finding the Parameters of Gaussians

In our data fusion approach each source of information is assigned with a two-dimensional Gaussian distribution that serves as an error model. This means that at each point in time the parameters of different distributions have to be estimated in order to combine them into one merged distribution using the method introduced in Section 3.1. In the case of WSN localization the estimated mean is the position computed by evaluating the measured distances. It can be shown that the measurement error is affected by the geometric properties of the network at the estimated position coordinates [Nie97]. We account for this by including the horizontal dilution of precision (HDOP) into the computation of the standard deviation of the WSN estimate. Thus, the WSN estimation errors along the coordinate axes can be written as:

$$\sigma_x^{WSN} = \sigma_s H_{(1,1)} \tag{8}$$

$$\sigma_y^{WSN} = \sigma_s H_{(2,2)}, \tag{9}$$

where σ_s is the assumed standard positioning error of the WSN and H is the matrix of position dependent dilution factors. The details of DOP computation determining H can be found in [Kap05]. Fig. 3a shows the resulting positional error estimate for different positions in a vehicle with $\sigma_s = 0.7$m. The geometry of the anchor nodes causes a higher error in horizontal direction of the WSN coordinate frame in this case. The effect increases with the distance to a central position in the network. Thus, it seems beneficial to include sensors that provide a good horizontal accuracy in the positioning process.

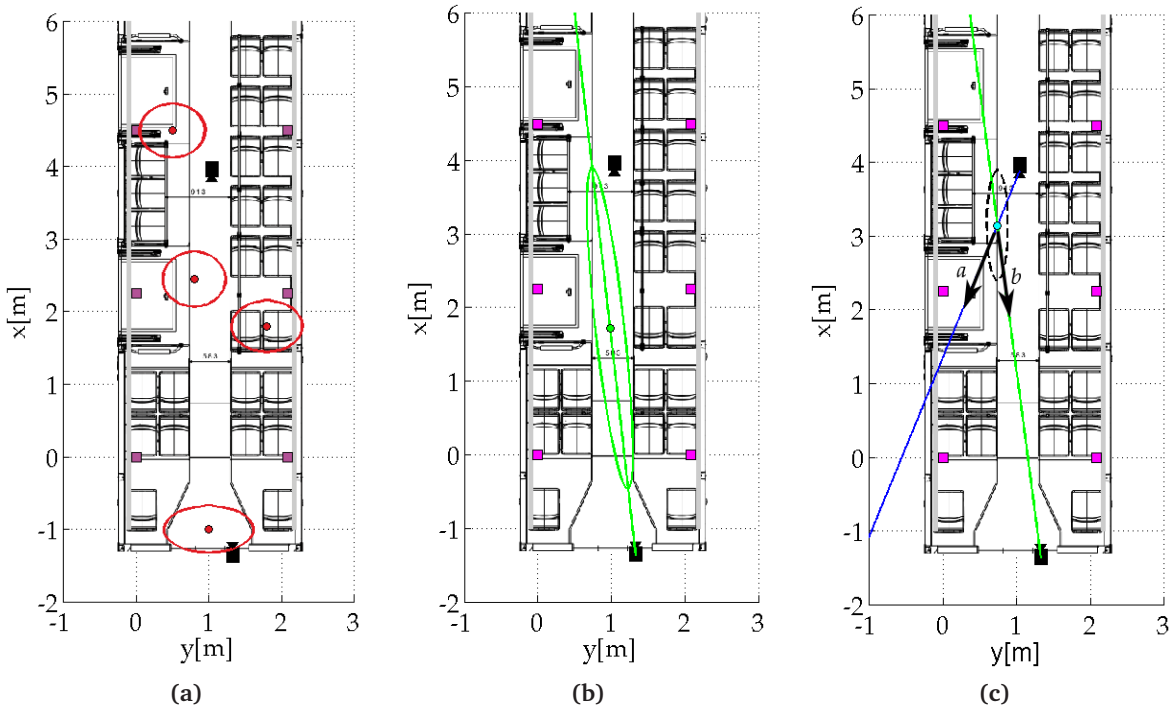


Figure 3: Sensor measurements in the WSN coordinate frame: (a) Estimated WSN positioning error using HDOP, (b) ray and error ellipse for an object detected in one camera view, (c) intersection of rays for the same object detected in two camera views

As pointed out in Section 2.2, an object visible in a camera image relates to a line in the WSN coordinate plane as there is not sufficient information from 2D image coordinates to determine the depth of an object. Thus, in order to model a 2D Gaussian for this kind of measurement we align its semi-major axis with the direction of the ray and assume $\sigma_{maj} \gg \sigma_{min}$ to account for the lack of depth information. Fig. 3b shows the resulting distribution. Its mean is assumed to be approximately the middle of the camera's detection zone in vertical direction.

In case two or more cameras detect the same object from different angles, its position in the WSN coordinate frame can be determined using the intersection of rays. The measurement error here relates to a inaccurate direction of the ray connecting camera center and the detected object. Thus, the error of this position measurement can be determined as a function

of the angle of intersection. The angle between two vectors can be written as:

$$\arccos \varphi = \frac{a.b}{|a||b|}, \tag{10}$$

where $a.b$ is the dot product of the vectors a and b, representing the direction of rays from the two cameras to the object in the WSN coordinate frame (see Fig. 3c). The values σ_{maj} and σ_{min} for the corresponding Gaussian distribution are then computed as:

$$\sigma_{maj} = \xi\varphi \tag{11}$$

$$\sigma_{min} = \xi(2\pi - \varphi), \tag{12}$$

for $\varphi \geqslant \pi$. The factor ξ scales the standard deviations to reasonable values while preserving the ration of angles. Fig. 3c shows the error ellipse resulting from detecting the same object in two camera views. It indicates a higher error in vertical direction of the WSN coordinate frame due to the geometry of the intersecting rays.

4 Experimental Results

Our data fusion approach takes input from two different cameras showing the same scene inside a public transport vehicle from different angles. In the test scene used for experiments a person walks a random path in this vehicle with a mobile sensor node in his right hand. A new position of the mobile node is computed approximately once per second based on the 10 latest distance measurements to the anchor nodes. Camera images are only retrieved when a new WSN-based position estimate is available. Measurements resulting from image evaluation are based on a state-of-the-art people detection algorithm [Dal05]. The result of a successful detection can be seen in Figures 4a, 5a and 5b. The red bar in these camera images connects the current WSN position estimate at a height of 0.8m and its projection on the vehicle's floor for reference. Both points are computed from 3D world coordinates using Eq. (1).

If no information is available from the cameras the final position estimate and its standard deviation in the WSN coordinate system are solely driven by the characteristics of the distance-based positioning method, thus the result does not differ from the examples given in Fig. 3a.

By contrast, Fig. 4 shows a scenario where a measurement from one camera is available. Combining the two 2D Gaussians that arise from the different sensors leads to an improved position estimate. Its mean is shifted towards the direction where the standard deviation of the camera measurement is assumed to be small, i.e. perpendicular the ray's direction. Contrary, we modeled the position of the detected person on the ray as highly uncertain. Thus, the component of the combined mean in this direction is mainly influenced by the WSN measurement, which features a much smaller standard deviation in this direction. Furthermore, the remaining uncertainty w.r.t. the WSN coordinate system is driven by a

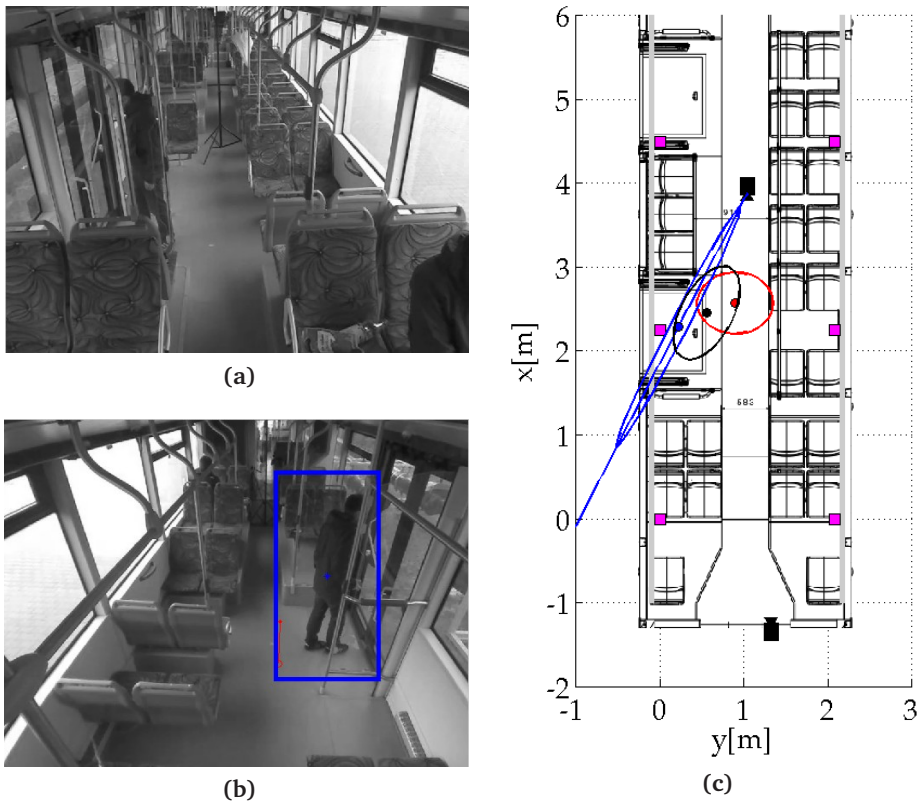


Figure 4: Sample scene with successful detection in one camera image: (a), (b) camera images with result of people detection (blue box) and WSN position (red bar), (c) WSN position estimat (red), camera-based position estimate (blue) and combined estimation (black) with corresponding error ellipses

mixture of the parameters σ_{maj}, σ_{min} and the angel θ of the individual distributions. This leads to the reasonable result that the combined standard deviation constitutes an ellipse that is aligned similar to the direction of the camera position estimate.

Fig. 5 showcases a scene where the person is detected in both camera images. Thus, a camera-based position estimate by computing the intersection of rays is possible. As described in Section 3.1, the lengths of the corresponding semi-major and semi-minor axes are a function of the angle of intersection. As in the previous case, the combined position estimate is a weighted mean of the two independent measurements. In both cases (i.e. measurements from one or two cameras available) the standard deviation in the direction of the y-axis of the WSN coordinate frame is reduced as a result of our data fusion method. This outcome is of particular interest as it helps improving the WSN position in the spatial dimension where an accurate and reliable position estimate is needed for advanced applications.

5 Conclusions and Future Work

In this paper we introduced a data fusion method for improving the localization of a person. This goal is achieved by combining two cameras and localization based on distance

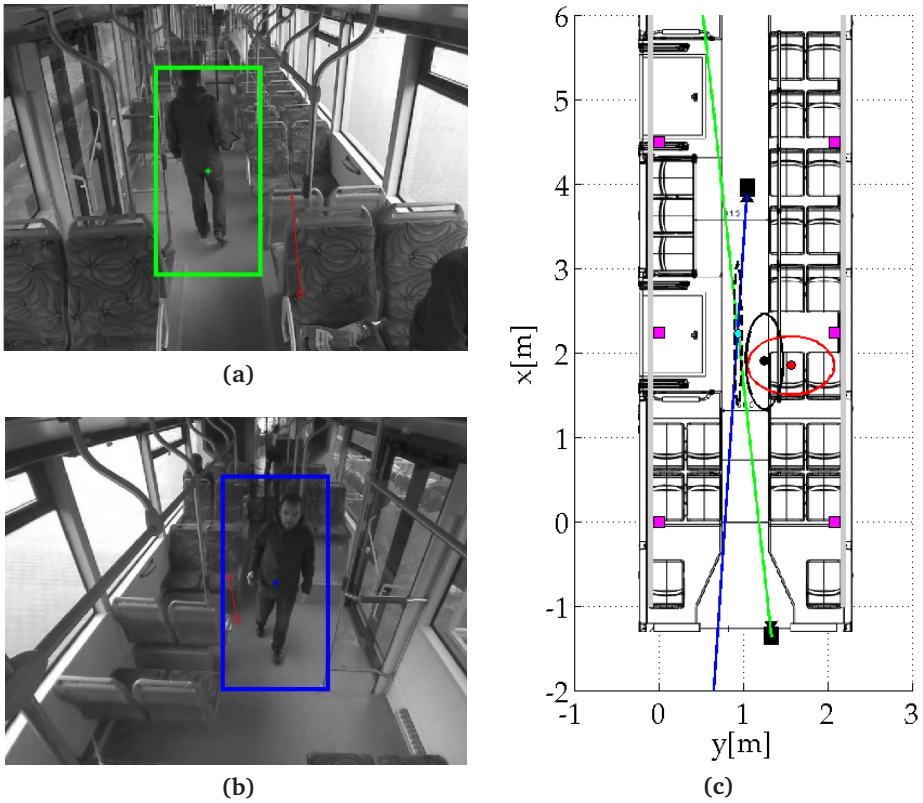


Figure 5: Sample scene with successful detection in both camera images: (a), (b) camera images with result of people detection (blue box) and WSN position (red bar), (c) WSN position estimate (red), intersection of camera rays and corresponding error ellipse (black dotted), result of combined estimation (black)

measurements acquired by a WSN. We showed how to model the positioning error of the individual sensors as 2D Gaussians in the WSN coordinate plane and how to accordingly combine them to one merged 2D Gaussian. This approach yields an improved positional mean and a better estimation of the standard deviation in many cases. Our algorithm summarizes the information available at one point in time and thus does not involve computationally intensive methods like continuous tracking of a mobile node. In the context of this work we simplified the problem significantly by restricting it to the localizing of only one person. In order to generalize our method to cope with a multitude of users the challenging problem of data association has first to be solved. That is, persons detected in different camera views and multiple mobile tags must be assigned to entities that represent the same passenger first. Furthermore, future work comprises the introduction of an improved camera-based detection algorithm that can also recognize partially occluded people.

Acknowledgment

This work was partly supported by the project Cool Public Transport Information (CPTI), a part of the leading-edge cluster COOL SILICON, co-funded by the European Union and the Free State of Saxony, Germany.

References

[Ben07] A. BENSKY: *Wireless Positioning Technologies and Applications*. Norwood, MA, USA: Artech House, Inc., 2007. ISBN: 1596931302, 9781596931305.

[Dal05] N. DALAL and B. TRIGGS: "Histograms of Oriented Gradients for Human Detection". In: *International Conference on Computer Vision & Pattern Recognition*. Ed. by C. SCHMID, S. SOATTO, and C. TOMASI. Vol. 2. INRIA Rhône-Alpes, June 2005, pp. 886–893. URL: http://lear.inrialpes.fr/pubs/2005/DT05.

[Har04] R. I. HARTLEY and A. ZISSERMAN: *Multiple View Geometry in Computer Vision*. 2nd ed. Cambridge University Press, 2004. ISBN: 0521540518.

[Kap05] E. KAPLAN: *Understanding GPS - Principles and applications*. 2nd edition. Artech House, Dec. 2005.

[Lan06] S. LANZISERA, D. LIN, and K. PISTER: "RF Time of Flight Ranging for Wireless Sensor Network Localization". In: *International Workshop on Intelligent Solutions in Embedded Systems*. June 2006, pp. 1–12. DOI: 10.1109/WISES.2006.329127.

[Nie97] R. NIELSEN: "Relationship between dilution of precision for point positioning and for relative positioning with GPS". In: *Aerospace and Electronic Systems, IEEE Transactions on* 33.1 (1997), pp. 333–338. ISSN: 0018-9251. DOI: 10.1109/7.570809.

[Str00] A. W. STROUPE, M. C. MARTIN, and T. R. BALCH: "Merging Gaussian Distributions for Object Localization in Multi-robot Systems." In: *ISER*. Ed. by D. RUS and S. SINGH. Vol. 271. Lecture Notes in Control and Information Sciences. Springer, 2000, pp. 343–352. ISBN: 3-540-42104-1. URL: http://dblp.uni-trier.de/db/conf/iser/iser2000.html#StroupeMB00.

[Web11] R. WEBER, R. RICHTER, O. MICHLER, and S. ZEISBERG: "RF-based positioning and localization techniques in wireless sensor networks using a C-MDS approach". In: *8th Workshop on Positioning Navigation and Communication (WPNC)*. 2011, pp. 111–115. DOI: 10.1109/WPNC.2011.5961025.

Corresponding author: Uwe Gosda, Technische Universität Dresden, "Friedrich List" Faculty of Transportation and Traffic Sciences, Institute of Traffic Telematics, 01062 Dresden, Germany, phone: +49 351 463 36755, e-mail: uwe.gosda@tu-dresden.de

Johnson Criteria applied for Traffic Incident Detection Systems

Johannes Traxler
TB-Traxler

Abstract

John B. Johnson did describe the image- and frequency domain approaches to determine the ability of observers to perform visual tasks using video signal sources. He was the first one who did develop a guideline to develop video systems based on their recognition requirements. Based on the "Johnson Criteria" many predictive models for image processing sensor systems have been developed to predict the performance under different environmental and operational conditions.

Nearly 60 years later this theoretical knowledge to design video based automatic incident detection systems (IDS) is not used in traffic surveillance applications. This caused disappointed customer because installations did not reach their expectations and a lot of efforts for all project partners. This paper reports a hypothesis of how to determine the target detection performance of traffic incident detection systems based on the "Johnson Criteria". It also suggests and explains a way of how traffic surveillance systems could be designed and tested.

Keywords: Jonson Criteria, ITS, Incident Detection Systems, Video Detection, Performance, Design, Testing

1 Introduction

Management of crossroad by traffic policeman may be seen as the earliest form of traffic surveillance. Television cameras and monitors have liberated us from the need to be up close to the monitored scene. Using CCTV-System operators are able to observe multiple remote scenes at the same time. This is one of the reasons why CCTV-System have spread fast through many areas of applications in industrial production, public and building safety and public and individual transport.

Video surveillance in traffic applications is a well-known and standardized technology. Because the video signals were already there, in the past decade many sites were extended with video based automatic incident detection systems (IDS). These installations are used to immediately detect any safety critical event. All the advantages of the IDS against common

sensor technology caused very high expectations of the customers. They started thinking about decreasing the amount of operators in the control rooms and building full automatic traffic surveillance systems. The manufacturer supported this expectations but nobody did take care on the technological limitations.

The result was on the one hand unfortunate users and on the other hand many different field studies. Second were not based on theoretical but on empirical results which did never reach the expectations of the customers about recognition rates (true positive) and false positive rates, where recognition means that the target object is detected and classified as that kind of object based on its dimension and shape. This is the reason why this paper presents a brief survey of the methods and procedure, which may be utilized in future for the designing of IDS in traffic surveillance.

2 General system consideration

Traffic surveillance systems with integrated IDS are using state-of-the-art camera technology. The Camera signals are transmitted over fibre optics or copper cable as analogue or digital signals to a control room. Depending on the kind of signal different kind of electronic equipment is receiving the video signals, encode, stream, store and analyse it. The streamed video signals are received in remote control rooms where operators are monitoring them or if necessary are accessing the video storages to play back past sequences.

In order to obtain usable semantic information in form of incidents, a processing chain of computer vision algorithms has to be applied. A sequence of digital images provided by the acquisition step is used as an input. Since video compression may negatively influence detection quality, algorithms operate commonly on raw digital images thus, placing detection close to image acquisition. When detection of different incidents is desired, a variety of techniques have to be applied for just one input camera. Three major categories of algorithms can be identified. The first group deals with events of a static nature such as lost goods obstructing driveways or monitoring break-down bays, while the second is in charge of dynamic behaviour, such as detecting wrong-way-drivers, traffic jams or stopping vehicles. The last group is a collection of highly specialized algorithms dealing with tasks like smoke recognition or the detection of persons.

Apart from the detection itself, another task has been added to digital CCTV-systems through the introduction of computer vision. When an incident has been identified through detection the surveillance system should respond in an appropriate manner. Usually the desired reaction includes automated creation of a recording and informing a human operator or an external security system about the incident through an alarm output. Therefore, the application has to be equipped with facilities to transport and deliver and visualize the alarm information. Therefore SCADA, short for Supervisory Control And Data Acquisition, is used.

In General traffic surveillance systems do have to fulfil the requirement to gapless observe the road section of interest. This means for automatic incident detection systems that the cameras have to overlap and in the border area objects (vehicle, pedestrian, …) have to be detected high reliable. These regions are the most difficult place to create high detection

rates (true positive) and low false alarm rates because all the limitations of resolution, lenses and signal-noise ratio have their biggest impact. But of course any kind of traffic incident detection system has to deal with this requirement. Therefore it is important to be aware of the consequences of increasing camera distances and focal length or image resolution for gapless incident detection systems.

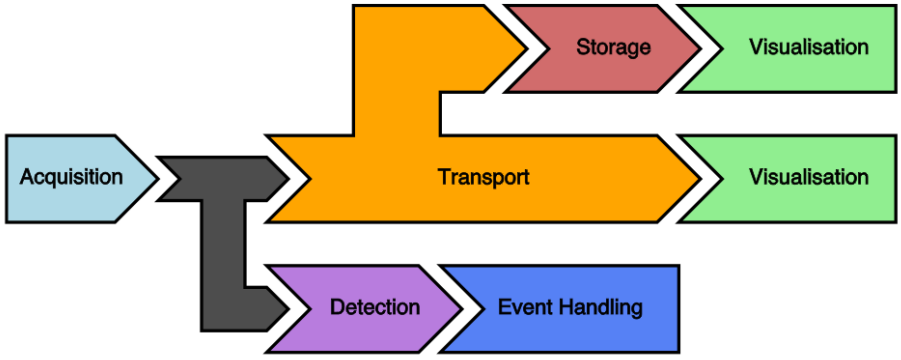


Figure 1: General Workflow for traffic surveillance systems.

3 Johnson Criteria

Like in each video driven application the quality of the systems depends very much on the resolution of the camera and the signal-noise-ratio (SNR) of the transmitted and received video signal. This general principal were researched by Johnson in the 1950´s because first applications mostly in military applications were growing up and the usability of such systems had to be analysed in theoretical manner [Joh58].

He found out that independent of what kind of detection algorithm or visual analysis you want to apply on the image there are always five distinct levels of activity:

1. No detection
2. Detection
3. Shape detection
4. Shape recognition
5. Detail recognition

These are the five distinct degrees of freedom or states of video systems. Obviously these decision states depend on the characteristics of the target object, the properties of the cameras and the quality of the detection algorithm or the physiological responses of the human readout process. To determine the value of decision he was looking for an arithmetic transformation. The result of this transformation would be the probable value of the decision state of the complete system as a function of the various components.

Therefore Johnson explained the dependency of contrast and brightness resolution with camera spatial resolution and scatter coefficient. Spatial resolution can be determined using a circular or rectangular object and measuring the response function of the camera according

to the diameter of the circular. The fact that as the image diameter d decreases below the dimension of the point response diameter d_r the peak amplitude of the output image function falls rapidly shows the loss of effective image resolution as a function of target distance or size.

For the theoretical description Johnson suggested the parameters:

$C_i = input\ target\ contrast$

$C_0 = output\ image\ contrast$

$R_S = camera\ spatial\ response$

$B_0 = output\ background\ brightness$

$B_0^T = output\ target\ brightness$

$B_i = input\ background\ brightness$

$B_i^T = input\ target\ brightness$

$B_0^N = output\ background\ noise$

$K = camera\ scatter\ coefficient$

He showed that the output image contrast C_0 is given by the equation:

$$C_0 = \frac{R_S(B_0^T - B_0)}{\frac{(B_0^T - B_0)}{C_i} + K(B_0 + B_0^T) + B_0^N} \tag{1}$$

while the output image brightness is given by:

$$B_0 = \frac{B_i^T TG}{4F^2M^2} + B_0^N + R_S\frac{B_i^T - B_i}{4F^2M^2} \tag{2}$$

3.1 Image analysis in space frequency domain

In space domain image analysis and evaluation is based on the output-input relationship indicated by

$$B_0(x^i, y^i) = \int_{-h/2}^{+h/2}\int_{-w/2}^{+w/2} B(x,y)f(x^i - x, y^i - y)dxdy \tag{3}$$

where

$h = height\ of\ the\ image$

$w = width\ of\ the\ image$

$B_0(x^i, y^i) = output\ image\ function$

$B(x, y) = input\ image\ function$

$f(x^i - x, y^i - y) = system\ point\ function$

It is obvious that image evaluation based on image in the space domain is a tedious operation that must be repeated for each view of each target of interest. Johnson suggested that the space frequency response method considerably simplifies the situation.

The equation for space frequency response is given by a Fourier-Transformation:

$$I(\omega) = \int_{-\infty}^{+\infty} B_0(x) \cdot e^{-i\omega(x)} dx \tag{4}$$

$I(\omega)$ is the output frequency spectrum while $B_0(x)$ is the output image function. However the resulting output image spectrum is given by

$$I(\omega) = O(\omega) \cdot f(\omega) \tag{5}$$

where

$O(\omega) = input\ image\ spectrum$

$f(\omega) = system\ frequency\ response$

3.2 Optical image transformation

Intuitively it would seem that there must be a dependency between the number of lines resolved at the target and the corresponding decision of detection, recognition or identification. Johnson found out that the resolution required for a particular decision activity was a constant for nine different target types. Because Johnson was working in military application he focused on tanks.

For the different types of tanks he got the following results about how many pair lines (interlaced) are required for the different decision levels:

Table 1: Johnson Criteria for different target types.

Broadside View	Detection	Recognition	Identification
Truck	0.90	4.5	8.0
M48 Tank	0.75	3.5	7.0
Stalin Tank	0.75	3.3	6.0
Centurion Tank	1.00	3.5	6.0
Jeep	1.00	4.0	6.0
Car	1.20	4.5	5.5
Soldier	1.50	3.8	8.0

These target transformation were found to be independent of the contrast of the scene and the signal to noise ratio as long as the contrast of the object was similar to the contrast of the background. These results indicated that targets with equivalent resolution patterns reach the same decision level with similar spatial frequency. Johnson concludes that the decision level activity in any imaging system is similar and it is only necessary to determine the angular characteristic as a function of a few parameters.

4 Traffic incident detection systems

The Johnson Criteria showed that independent of the kind of imaging system the decision states can be determined according to the image and frequency representation of the target objects (see section 3). Johnson uses the resolving power of an imager as a metric of sensor "goodness" for target acquisition purposes [Joh58]. For a given target to scene contrast, resolving power is the highest spatial frequency passed by the sensor and display and visible to the observer [Lea03]. He multiplies the resolving power of the imager (in cycles per milliradian) by the target size (in milliradians) to get "cycles on target".

It is required to assume some principal settings for the cameras and the video system, so that following calculations can be done. First we assume that for all the following calculations the viewing angle of the camera is the half of the opening angle of the lens. Second we do not take care about light situation and camera sensor dynamic characteristics. This would require additional analyses that were not done here. But it has to be clear that this characteristics influence the results too and for the system design they have to be considered. Additionally we assume that the camera is producing analogue video signal in PAL or NTSC standard.

4.1 How to apply Johnson criteria described on traffic incident detection systems?

In traffic applications video based incident detection systems do have some general conditions coming from the kind of objects and movement. Primarily vehicles are the target objects to be detected. Of course also pedestrian, animals and cargo are important but this kind of target objects may be analysed afterwards. So coming back to the vehicles they have the characteristic to move in a linear manner. In other words this kind of object are not able to change their moving direction within a short time period. Also they are not changing their shape during their movement through the observation zone.

This both important conditions of vehicles are the basis for tracking algorithms in traffic incident detection systems. It is not possible to explain all the possible solutions here, but they are well known and can be read in the "Wissenpapier" [Vid13]. All of the different kind of algorithms does need to detect as a minimum the incident types:

- Wrong way driver,
- Slow vehicle,
- Stopped vehicle,
- Traffic jam.

The event types wrong way driver, slow vehicle and traffic jam are not position sensitive. Objects that are driving in the wrong direction or slow do move during the observation phase. So they enter the detection zone of the specific camera and move through it. So independent of they are moving towards or from the camera there will be a situation where

the object is occupying so many lines of the image that compared with the Johnson Criteria the level of shape recognition will be reached. Of course it will be necessary to do so in more than one frame because it is necessary to determine the speed of the object, but this could be done in the part of the image where the representation object size is at the maximum. In other words this detections are not influenced by the Johnson Criteria very much.

Stopped vehicle, pedestrian, animals and lost cargo are left. These kinds of objects are of course different and do have different characteristics. All of them can appear on each position in the image or on the road and do not move. So different to the other incident types described above they have to be detected also at the border of the detection zones. This is the reason why following section are focusing on object sizes in worst case situations and do not take care on common situations. It is important for each incident detection system to fulfil the true positive rates not only in the near field of the camera but everywhere.

4.2 Resolution depth

Using the optical image transformation of Johnson (see 3.2) the image can be transformed onto the road layer. Each milliradiant of the camera image represents a specific length on the road. At the far field of the road one miliradiant is representing a longer distance as in the near field. That means that objects situated in the far field occupy less milliradians and as a consequence less image lines than in the near field, what is quiet obviously.

But how to determine the resolution limits for such an imager system? Similar to Johnson the minimum cycles on target do define the maximum detection distance. Obviously there is a dependency of focal length, viewing angle, mounting position and line resolution of the camera and the cycles on the target. Johnson did determine the minimum lines on the target for military targets.

So the results from Johnson given in Table 1 can be used for vehicles as follows:

Table 2: Results for different target types

Rear View	Detection	Recognition	Identification
Truck	0.90	4.5	8.0
Car	1.20	4.5	5.5
Pedestrian	1.50	3.8	8.0
Cargo	1.70	4.3	6.5

In Table 2 the minimum amount of lines, i.e. cycles on the target, is shown. Each line is in real represented by two lines in the full frame. But because of the blur effect usual only half size of the image is used for automatic incident detection. These values are the physical limits where a human is able to reach the different level of activity.

To get an idea of how objects do look like under this conditions the Fig. 2 – 4 show a car with the resolution for detection, recognition and identification with different line representations.

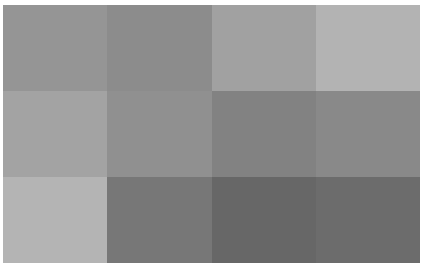


Figure 2: Car image representation at level of activity “Detection” with 2 lines.

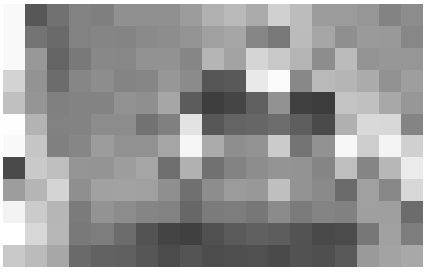


Figure 3: Car image representation at level of activity “Shape detection” with 6 lines.

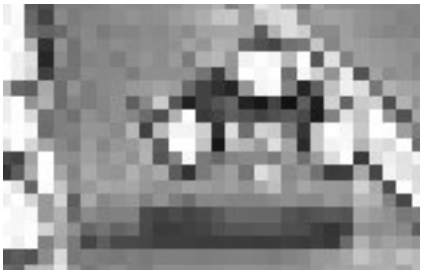


Figure 4: Car image representation at level of activity “Shape recognition” with 12 lines.

With level of activity „Detection“ it is for detection algorithm impossible to differentiate between a noise in the image or an object. Even the level of „Shape detection“ with a resolution of 6 lines (see Fig. 2) makes the detection algorithm unreliable. So for a real reliable and object type sensitive incident detection level of „Shape recognition“ with more than 12 lines is required.

This means in the manner Johnson described his optical image transformation (see 3.2), that in the worst distance from the camera the detection algorithm requires this resolution. So if a gapless incident detection system is required it is necessary to reach this line resolution of at minimum 12 lines at each place on the observed road. Based on that information the following equation gives minimum angle dependent of the used focal length and camera chip dimension, assuming an ideal lens:

$$\alpha_{min} = 2 \cdot arctan\left(\frac{m_{lines} \cdot \frac{d_c}{2}}{a_r \cdot 2 \cdot r_c \cdot f_l}\right) \tag{6}$$

where

$a_r = aspect\ ratio\ (common\ 4{:}3\ or\ 16{:}9, here\ 4{:}3\ is\ used)$

$d_c = camera\ chip\ diagonal$

$r_c = camera\ chip\ line\ resolution$

$f_l = focal\ length$

Using a standard analogue camera with PAL resolution a chip diagonal of ½" and a focal length of 18 mm equation 6 gives $\alpha_{min} = 22\ mrad$. Based on that result and the equation

$$L \doteq \frac{O_h}{\tan(\frac{\alpha_{min}}{2})} \tag{7}$$

where

$L = maximum\ vehicle\ distance,$

$O_h = vehicle\ height,$

it is possible to calculate the maximum distance for the object to be detected in a reliable manner. If we assume the average height of a car with 1.4 m the maximum distance to detect the object using automatic algorithm is 127.3 m. Of course this is the theoretical maximum distance where a car can be detected if the signal-noise ratio is low and the contrast of the image is perfect. In Fig. 4 an example of a usual car in a distance of 127m is shown. Of course it is for a human easy to recognize that there is a car but for a mathematical algorithm it is on the theoretical limit.

The conclusion of the results given above is, that the design of the video based incident detection system has to be done based on the required target object line representation in the video images. All system design efforts have to lean on this core requirement.

Figure 5: Car image in a distance of 127m.

4.3 Influence of focal Length

The quality of Video based Traffic incident detection systems can be described by two criteria. The sensitivity or true positive rate and the specifity or false negative rate. Both values depend of course very much on the object representation inside the image or in the words of Johnson the amount of lines that are captured by the object. Because object area inside the image is decreasing with increasing distance from the camera the worst-case situation that must be dealt with is at the end of the detection zone when the object is already entering the detection zone of the next camera.

Because nowadays still most of the traffic incident detection systems are using analogue video cameras the maximum detection depth for cars is 127,3m with a focal length of 18mm (see 4.2). This is the theoretical maximum distance if the signal noise ration is perfect (above 60 dB) and the contrasts are well. How is it possible to use these results for the designing of video based incident detection systems?

On the one hand it is obviously that the maximum working depth of the incident detection system influences the amount of camera in the way that for straight roads the length of the road divided by the maximum depth

$$N_{cam} = \frac{L_R}{\mathrm{L}} \tag{7}$$

where

$L_R = length\ of\ the\ road\ to\ be\ observed$

It is natural that if you want to detect other kind of objects like pedestrian or cargo with other object height O_h (see eq. 7) decrease L what leads to a higher amount of cameras, if resolution or focal length can not be increased.

Focal length is not a one-way solution because you do improve the detection depth but you also increase the distance where the image shows the complete road. This is the first position objects can be detected from the following camera.

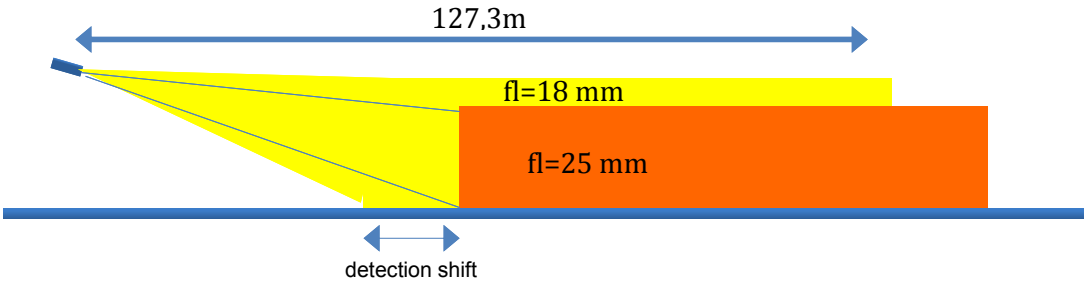


Figure 6: Focal length influence for the field of view.

In Fig. 6 the detection shift is shown between a 18mm and a 25 mm lens. Obviously occlusion is getting worth if the viewing angle because of the focal length is getting flat. Also the measurement of the moving speed of vehicles determined by the pixel movement of the object is less accurate if the viewing angle is lower. Not only the theoretical resolution limitations but also occlusion, moving speed measurement and deviated values are

influenced by the focal length and viewing angle. It is important to recognize all these influences when designing an automatic incident detection system.

4.4 Speed measurement for stopped and slow objects

The moving speed is especially important for the incident type stopped and slow vehicle. At the end of the detection zones the objects are smallest, i.e. the measurement of the movement is not accurate. This is influenced by the viewing angle of the camera the dispersion δ_v for the velocity is:

$$\delta_v = \frac{\mathrm{h} \cdot \tan(\arctan\left(\frac{L}{h}\right) + 2 \cdot \arctan(\frac{d_c \cdot a_r^{-1}}{2 \cdot f_l \cdot S_L})) - L}{t_{fps}} \tag{7}$$

where

$a_r = aspect\ ratio\ (common\ 4{:}3\ or\ 16{:}9, here\ 4{:}3\ is\ used)$

$L = vehicle\ distance$

$h = mounting\ height\ of\ the\ camera$

$S_L = amount\ of\ lines\ of\ the\ camera\ image\ sensor\ (f.e.\ 288\ or\ 640)$

$f_l = focal\ length\ of\ the\ lenses$

$t_{fps} = timeperiod\ according\ to\ the\ frames\ per\ second$

$d_c = camera\ sensor\ chip\ diagonal$

For example for a mounting height of $h = 4.5$m with a focal length of $f_l = 18$ mm, S_L=288 (non interlaced lines) and at a Distance of 127.3 m the dispersion of the speed is $\delta_v =$ 45,89 km/h. This appears to be quiet high but if you imagine that one image line is representing 6.6 m and using $t_{fps} = 40\ msec$ it results in such a huge dispersion if the object tracking only fails for one line caused by noise.

Table 3: Results for different target types for speed measurement.

Rear View	Recognition	Incident Detection
Truck	14.5	16.3
Car	12.0	14.0
Pedestrian	8.0	8.0
Cargo	9.0	12.0

5 Performance requirements

Considering the physical limitation described in section 4.1 – 4.4 about resolution depth, focal length and speed measurement the results based on the Johnson Criteria in Table 2 have to be reworked. The limitation of the speed measurement, described in section 4.4, obviously the detection depth based on the object resolution can not be reached for the

incident types stopped vehicle, slow vehicle, pedestrian or lost cargo. The results for common traffic incident detection systems then are:

Please consider that all these values have been calculated for a straight road. Special road features like curves influences the detection depth. In any case the minimum amount of lines described in Table 3 is valid but perspective distortion may have bad impact. So these kinds of environmental conditions require specific effort to get an optimum system design and cannot be described general.

6 Concluding remarks

Using the Johnson Criteria it has been presented that the resolution performance of the video camera system is essential for planning and designing of video based traffic incident detection systems. The minimum amount of lines representing an object type gives the design according to distance and mounting position. It was shown that in addition to the Johnson Criteria motion determination based on object tracking results is limited. This causes an additional limitation of detection depth for special kind of detection types like stopped vehicle or slow vehicle. Of course all these methods are approximations that require refinement for the conditions of the special site.

References

[Lea03] J. LEACHTENAUER: "Resolution requirements and the Johnson criteria revisited". In: *Infrared Imaging Systems: Design, Analysis, Modeling, and Testing XIV*. Vol. 5076. Aug. 25, 2003. DOI: 10.1117/12.

[Joh58] J. B. JOHNSON: "Analysis of Image Forming Systems". In: *Image Intensifier Symposium.* U.S. Army Engineer Research and Development Laboratories, Ft. Belvoir, VA, Oct. 6-7, 1958.

[Vid13] ARBEITSKREIS VIDEO: *Wissensdokument Videodetektion*. Forschungsgesellschaft für Straßenverkehr, Köln, Sept. 2013.

Corresponding author: Johannes Traxler, TB-Traxler e.U., Pfeningberg 23, 3500 Imbach, Austria, phone: +43 664 730 14 461, e-mail: Johannes@tb-traxler.at

Vehicle Tracking using 3D Particle Filter in Tunnel Surveillance and Incident Detection

Adrian Fazekas, Michael Bommes, Markus Oeser
RWTH Aachen University

Abstract

Traffic surveillance aims at detecting incidents and accidents to provide prompt and appropriate reaction. Tunnels deserve special safety concerns, since accidents in tunnels occur less often than outwards, but include a highly increased risk potential. Automatized video surveillance is preferable, since humans suffer from operator fatigue on long term surveillance tasks. Tracking can be used to understand individual traffic participant's behaviour, allowing the extraction of stable properties such as vehicle velocity and trajectory in order to forecast incidents. The research presented in this paper shows how particle filtering can be used to enable vehicle tracking in traffic surveillance tasks. The presented method is qualitatively evaluated for a recorded real tunnel traffic data use case, while quantitative run-time performance analysis shows real-time capability on standard hardware.

Keywords: tracking, video surveillance, tunnel, particle filter, condensation, camera calibration, cpu, real-time, coarse 3D models

1 Introduction

In the past years, intelligent transportation systems (ITS) have been an area of active research as the continuous increase of on-road traffic not only leads to an inefficient usage and degradation of present infrastructure but also induces additional risk of accidents. Incidents in tunnels require thorough considerations due to the limited space and hazardous environment in case of fires. This leads to the necessity of developing automated tunnel surveillance systems which help tunnel operators to instantly detect incidents and react accordingly. Different sensor technologies can improve traffic surveillance and deliver necessary data for roadside traffic and incident management. Notably, inductive loops are widely used and reach high measurement accuracies but also suffer some considerable drawbacks. On one hand, ground loop detectors are inconvenient and costly to install and to maintain as implementation requires excavation of road surface. On the other hand, as they are spot monitoring sensors they can only deliver a limited amount of data such as vehicle count and speed. Incident

detection methods are based on this few traffic parameters and this technology often has long detection delay times [Lan12]. In contrast, video sensor technology is cheap and easy to install and to maintain. Furthermore a video-based traffic monitoring which constitutes an area monitoring system is able to directly detect various incidents rather then indirectly conclude from other measured traffic parameters. Video based detection has come more and more into focus of research while the performance of computers has risen to a level where sophisticated image processing algorithms can be used in real-time traffic monitoring systems [Kim13].

Although some research has been done on the topic of automated traffic monitoring, at present many deployed systems have algorithmic limitations due to the necessity of real-time application. Some of the work surveyed in [Buc11] was based on detection techniques through background modelling which tend to suffer in accuracy from shadows, change of direction and size in perspective and occlusion. To overcome these difficulties authors have proposed algorithms including 3D modelling of vehicles combined with perspective calibration of the camera, a technique introduced in [Kol93]. In [Buc10] a method was proposed, which combined background estimation with the 3D model comparison, for initial detection and classification of road users. Another approach to vehicle classification was presented in [Hod12] where initial detection was realised through a general object detector [Fel10] whilst for classification and pose detection edge based methods were used.

A robust and prompt incident detection requires stable tracking method of road users. Approaches using tracking-by-detection tend to have low accuracy due to shortcomings at the detection step in image areas where the size of the perspective projection of the tracked object lessens. In [Haa99] a Kalman Filter method is used for tracking, which is able to deliver prognosis on the movement of objects. Similar technique with enhanced flexibility on the prediction was employed in [Zha10] by using particle filters. State of the art detection systems combine edge based detectors with additional feature extraction methods, include 3D modelling of the scene and vehicles whilst tracking by prediction.

In this work we present the use of particle filters with three-dimensional state vectors. Camera calibration represents the basis of the tracking technique, thus we will give a short description of the this method in section 2.1. We will continue on to describe general probabilistic tracking techniques in section 2.2 and the theory behind discrete representation of distributions in section 2.3. In section 2.4 the state spaces used will be presented while section 2.5 will introduce particle filtering as a specialisation of Bayesian filters. The following subsection will introduce the similarity measure used in the correction step, after which implementation details of a specific application will be presented in section 2.7. Section 3 contains a summary of the result and a discussion including ideas for future work will finalize the paper.

2 Methods

2.1 3D calibration

The calibration of the scene is done by using image features on the road surface like edges of markings. These feature points have a known relative position either through the expectation of normed road marking lengths and lane widths, or through exact measurement of strong edge points in the scene. By the knowledge of enough points on the road surface and their projected pixel coordinate in the image a fast calibration can be performed using [Zha00]. This gives the possibility to project each scene point to the image plane. Underlying transformation for the calibration is homography which translates two dimensional points from one plane to the other and vice-versa. In this case we make the assumption that the road is plane and constitutes one of the two planes of the homography-transformation, the other being the image plane. This gives us the possibility to make an inverse projection of points in the image plane to the road surface, which is needed in the initialization phase of our work.

2.2 Bayesian tracking

In vehicle tracking tasks, we are interested in extracting object motion information from a sequence of frames to infer information about driver behaviour or traffic incidents.

Simple tracking methods include *tracking by detection* and *tracking using matching* [For12]. In *tracking by detection*, object candidates are detected in each frame by an appropriate detector and linked to previous frames' objects by examining spatial relations. In *tracking using matching* tracked objects' candidates will be searched in a region of prediction and compared to the original object, until a convenient candidate is found. *Bayesian tracking* instead is an approach considering likelihoods over the whole time sequence, enabling more plausible results with less dead ends.

In Bayesian tracking the *hidden state* $x_k \in \mathbb{R}^d$ represents position and other relevant properties of the tracked object described by random variable X_k, regarding the *k*-th frame of a video stream. The *measurement/observation* $y_k \in \mathbb{R}^c$ is the observable state described by random variable Y_k. Random variable histories we will abbreviate as $R_{i:j} = [R_i, R_{i+1}, \ldots, R_j]$ of random variables R. In Bayesian tracking the aim is to estimate the posterior probability $p(X_k|Y_{0:k})$ which is the distribution of object states at frame number k given the complete history of measurements. Given the assumptions, that the current measurement/observation is independent from previous measures and states (1) and that a new real state is only influenced by the last real state (2) (first-order Markov chain model). The posterior probability distribution can be rewritten as shown in (3) using these assumptions, Bayes rule and marginalization, while the constant $\kappa = \int P(Y_k|X_k) \cdot P(X_k|Y_{k-1}) \, \mathrm{d}X_k$ is used for normalization to receive a distribution.

$$P(Y_k|X_k, Y_{0:k-1}) = P(Y_k, X_k) \tag{1}$$

$$P(X_k|X_{k-1}, Y_{0:k-1}) = P(X_k|X_{k-1}) \tag{2}$$

$$P(X_k|Y_{0:k}) = P(Y_k|X_k) \cdot \int P(X_k|X_{k-1}) \cdot P(X_{k-1}|Y_{0:k-1}) \ \mathrm{d}X_{k-1} \tag{3}$$

The integral holds the posterior probability $P(X_{k-1}|Y_{0:k-1})$ of the previous frame and the probability of the new state knowing the old state $P(X_k|X_{k-1})$, which will be evaluated by a model describing the object movement/dynamics (sometimes called the *temporal prior*). The observation likelihood $P(Y_k|X_k)$ can be computed by knowledge about the object appearance in the video frame for a given state X_k. Depending on the previous states posterior probability distribution, this clearly is a recursive computation of posterior probabilities.

To extract the most likely object state, the posterior probability density function finally has to be evaluated by an appropriate method.

2.3 Weighted sample representations of distributions

Often posterior probability is a multi-modal distribution, which makes it hard to be represented and evaluated. One way of representation is to use weighted samples as a non-parametric approximation. In principle, any probability distribution $P(U)$ can be represented by a discrete set of N weighted samples [For12] $M = \{x^{(i)}, w^{(i)}\}$ with $|M| = N$ which approximates the distribution. This is, the distribution is sampled in regular or irregular intervals. The weights are computed relative to their function value and the probability in which the corresponding sample was chosen. Hence for regular intervals weights are chosen proportional to the function values. Irregular intervals aim to represent high density function values with more samples, since these values contribute more to applications on distributions like computation of expectation values. On the other hand, regions represented by fewer samples need their weights to be increased to approximate the real distribution. So samples $m^{(i)} = \{x^{(i)}, w^{(i)}\}$ consist of states $x^{(i)}$ drawn by distribution $S(X)$ (with density function $s(x)$) and their corresponding weights $w^{(i)} = \frac{f(x^{(i)})}{s(x^{(i)})}$ for density function $f(X)$ of represented distribution.

To approximate the original distribution, a Gaussian mixture component can be constructed for each sample $\{x^{(i)}, w^{(i)}\}$, building a mixture distribution $\tilde{p}(v)$ which can be evaluated at each value v as shown in (4).

$$\tilde{p}(v) = \sum w^{(i)} \mathcal{N}_{x^{(i)},\sigma^2}(v) \tag{4}$$

However the samples are mostly used directly to evaluate the integral by Monte Carlo sampling methods.

2.4 State space and temporal dynamics

In visual tracking applications, the state includes some representation of the object position within the image, together with all relative or interesting information. General state space models include the 2D image position of an object and some scale information (uni-

variate scale factor or perspective deformation), possibly enhanced by some velocity and/or acceleration information for importance sampling.

In specialised applications, state space can be adjusted according to the understanding of the object behaviour and properties in the real world.

We have chosen a state consisting of a 2D object position on the street embedded in 3D space, assuming that the street is planar, and a 2D velocity component, describing the position change per time on the street. We assume vehicle i to be a rigid object of constant size (length, width, height). In addition we assume it to be aligned basically along the street, so no rotations are used to describe the alignment more precisely. This way, a state $x_{t,i}$ of vehicle i at time t is described by its embedded position $p_{3D} = (p_x, p_y, 0)$ and its velocity (v_x, v_y), with relevant information combined as in equation (5).

$$x_{t,i} = (p_{x,t,i}, p_{y,t,i}, v_{x,t,i}, v_{y,t,i}) \tag{5}$$

To evaluate the visual information corresponding to that state, a 3D vehicle extent is computed from the embedded 3D position and the constant vehicle size information. This representation is projected to the camera image plane, where its extent defines the image region corresponding to the vehicle state. On the other side, given an image pixel, the corresponding

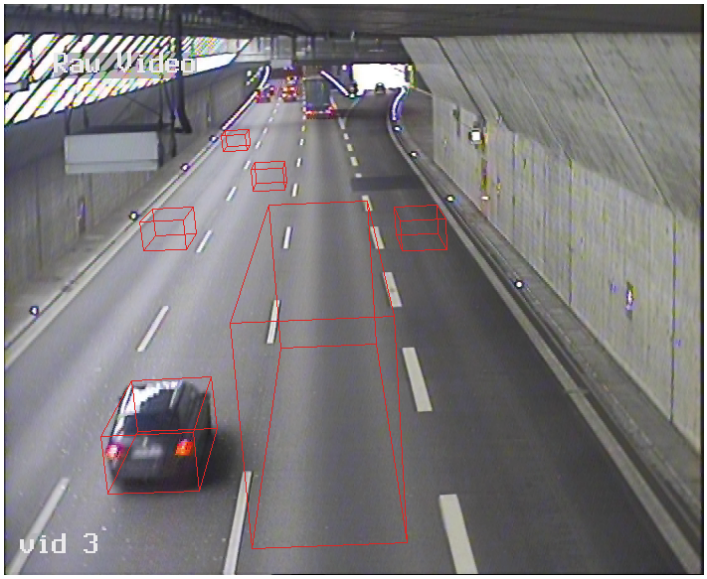


Figure 1: Vehicle states – This image shows a typical tunnel camera view. The red projections of 3D boxes mark some sample vehicle position states: Five passenger cars of equal size and a large goods vehicle. Comparison to the image of a real car shows the similarity to the state sizes.

point on the 3D street plane can be determined with help of the camera the calibration, so it is possible to convert an object state to the corresponding image region and vice-versa. In figure 1 example state projections are shown for two different vehicle sizes.

Since we describe the objects in scene space, the temporal dynamics describes object motion in scene domain. Object motion is evaluated by a first-order autoregressive model

where velocity is modified by a random acceleration, while the new position is determined by the sum of old position and velocity.

$$x_{t+1,i} = (p_{x,t+1,i}, p_{y,t+1,i}, v_{x,t+1,i}, v_{y,t+1,i}) \tag{6}$$

$$v_{x,t+1,i} := v_{x,t,i} + \mathcal{N}(0, \sigma_x^2) \tag{7}$$

$$v_{y,t+1,i} := v_{y,t,i} + \mathcal{N}(0, \sigma_y^2) \tag{8}$$

$$p_{x,t+1,i} := p_{x,t,i} + v_{x,t+1,i} \tag{9}$$

$$p_{y,t+1,i} := p_{y,t,i} + v_{y,t+1,i} \tag{10}$$

In scene domain temporal dynamics for special object classes it might be possible to use background knowledge to limit the random acceleration to known constraints[1]. In addition, scene domain states allows us to extract object dynamics from measured object positions.

2.5 Particle filter

Particle filtering is an implementation of recursive Bayesian filter by Monte-Carlo simulations, which is a Sequential Monte Carlo method. In particle filtering, the demanded posterior probability (3) is represented by a set of samples and weights which are called particles. To compute the posterior probability distribution of the next time step, particles are propagated according to a temporal dynamics model approximating the possible objection motion to get a prior. The outcome is modified by the likelihood of the measured image displaying the object with particles' new states, which again is a sampled distribution considered weight adjustement.

In Bayesian tracking context, it has been shown that application of temporal dynamics to each samples' state and adjustment of each samples' weight according to the likelihood of the prior results in the correct posterior probability distribution represented by those new samples [For12, pp. 385-391].

In detail the following steps have to be performed to track vehicle number i first appearing in time t:

Initialization: Samples have to be drawn to represent the first state distribution of the object. A set of samples $M = \{(x_{t,i}^k, w_{t,i}^k)\}$ has to be randomly selected to represent $P(X_{t,i})$ for $k \in \{0, \ldots, N-1\}$, where the weights are chosen as described in 2.3.

Prediction: In the prediction step $P(X_i|Y_{0:i-1})$ is represented by applying temporal dynamics to samples of $P(X_{i-1}|Y_{0:i-1})$. To represent the prior for time step $t+1$, all particles are moved according to the temporal dynamics model as shown in 2.4 while the weights are adopted unaffectedly.

Correction: Represent the posterior probability $P(X_i|Y_{0:i})$ using prior and likelihood. To

[1] i.e. the fastest street legal vehicles' acceleration at present is approximately $12.1 m/s^2$.

compute the posterior probability, the likelihood of the object being at a concrete position is computed for each particle position. Multiplying the weight of a prior with the likelihood results in a sample for the posterior probability distribution representation.

Resampling: Represent the same posterior probability by a different set of samples to reduce variance of weights. Therefore the wanted number of samples is drawn according to the distribution built by the weights.

In detail these steps are explained in the literature [For12, pp. 385-391]. The number of particles that have to be used to expect good results depends on the dimensionality of the state space, the accuracy of the motion model and the expected number of peaks of the likelihood.

The assumed object state in time step $t+1$ can now be approximated as the expected value of the sampled posterior $E_{posterior} = \frac{1}{N}\sum_{i=0}^{N-1} w_{t+1,i} \cdot x_{t+1,i}$ or the maximum a posteriori approximated by the maximum weighted particle $E_{map} = \max_{w_i}(x_{t+1,i})$

2.6 Similarity measure

In Bayesian tracking (and therefore particle filtering), the likelihood $P(Y_k|X_k)$ has to be evaluated, which is the likelihood of the measured data to represent a vehicle in state X_k. From the image region corresponding to that state, a similarity to the tracked object has to be computed if a high similarity indicates a high likelihood. In literature, comparison of color histograms are often used to measure similarity because of its invariance to some object transformations and its relationship to probability distributions. We use *sum-of-absolute-differences* (SAD) of image intensities as shown in (11) given two grayscale image regions $A, B \in \{0, \ldots, 255\}^{w \times h}$ of equal size.

$$SAD(A,B) = \sum_{j=0}^{w-1}\sum_{i=0}^{h-1} |A_{i,j} - B_{i,j}| \tag{11}$$

To estimate the likelihood from a given sum-of-absolute-differences, it is useful to know that finding the corresponding image area with minimal SAD is equivalent to a maximum likelihood estimator (ML estimator) with the assumption that remaining differences follow independent Laplacian distributions [Pat07].
The density function $f(x, \mu, \sigma)$ of the Laplacian distribution with mean μ and variance $2\sigma^2$

$$f(x,\mu,\sigma) = \frac{1}{2\sigma} e^{-\frac{|x-\mu|}{\sigma}} \tag{12}$$

is used as the likelihood of two image regions A, B representing the same object for $\mu = 0$, $x = SAD(A,B)$ and a fixed σ. Patras et al. choose σ depending on the variance of image region pixels intensity [Pat07].

The model is chosen as the initial visual representation of the tracked object in the image, either chosen manually by a user or extracted by an object detector.

Figure 2: Particle seeding – The images show particle seeding of the current frames, active particle state positions are displayed in white, the maximum likelihood particle is shown in black. on the left, the large goods vehicle is initialized, leading to a bigger number of particles and an increased standard deviation while the image on the right shows the situation 3 frames ($\approx 120ms$) later with a low variance set of particles.

2.7 Implementation details

Vehicles expected starting velocities are initialized to $v_{x,start} = 90km/h$. The number of particles is fixed to 30 per tracked vehicle, but to find an initialization, 90 particles are seeded with a standard deviation equalling the start velocity, such that about 70% of the particles will be seeded with guessed velocities within $[0, \ldots, 180]km/h$, regarding very most of all vehicles. Impressions of particle seeding for initialization and basic iterations are shown in figure 2. Real-time performance is critical for surveillance tasks to achieve fastest response times. Since the complexity of the method grows linearly in the number of particles and the number of particles is fixed per vehicle, we are interested in the time needed to perform one tracking iteration of a single vehicle. We have implemented the particle filter using task parallelism on CPU, allowing to perform tracking for different vehicles in parallel. We resize the image regions used in SAD similarity measure to achieve a constant SAD processing time and to compare equal sized regions unaffected by the difference of projective foreshortening. The similarity is computed on grayscale images, since we noticed a strong color noise in our image data. At any time t the assumed object state can be evaluated as being the maximum a posteriori of the sampled posterior distribution. We have chosen to resample the particles in each iteration.

3 Results

For evaluation we had access to a 8:19 minutes real world tunnel surveillance video sequence consisting of 12489 single images recorded with a frame rate of $25Hz$. This sequence contains

250 vehicles leading to a moderate traffic volume. Evaluating the video sequence using a preliminary detection module, 240 vehicles could be detected, but 5 of them were detected inaccurately resulting in initial states covering less than 50% of the vehicle visible in the image. Considering the other 235 vehicles correct tracking of 97.44% was achieved, where a vehicle is tracked correctly if more than 50% of the vehicle image is covered by the projected state in every time step until a maximum distance to the camera is reached. The maximum distance was chosen empirically, at a point where the perspective foreshortening becomes too big, resulting in tracking a vehicle for about 85 meters. Our implementation was tested on an Quad-Core Intel(R) Core(TM) i7 CPU 920 @ 2.67GHz using OpenMP task parallelism over vehicle representations. Since the number of vehicle and hence the number of particles varies over the sequence, the mean value of a single vehicle tracked was computed over all images where at least one vehicle occurs, leading to an average computation time of 2.02 ms tracking time per frame. The standard deviation of 1.12 ms appears to the costly initialization phase and the influence of interpolating image regions of larger vehicles during scaling.

Discussion and future work

While high tracking success rate is achieved, our method seems to be real-time capable for a fair amount of concurrently tracked vehicles. To increase the tracking performance, the used likelihood estimation might not be distinguishable enough, especially regarding low structured vehicles like some large good vehicles. In future work we would like to combine different similarity measures to build a more sophisticated likelihood estimation.

Acknowledgement

The work reported in this paper is part of the joint research project ESIMAS which is funded by the German Federal Ministry of Economics and Technology (BMWi) and supported by the project executing organization Mobility and Transport Technologies of the TÜV Rheinland Group (PT MVt). The support is gratefully acknowledged by the authors.

References

[Buc10] N. BUCH, J. ORWELL, and S. VELASTIN: "Urban road user detection and classification using 3D wire frame models". In: *Computer Vision, IET* 4.2 (2010).

[Buc11] N. BUCH, S. VELASTIN, and J. ORWELL: "A Review of Computer Vision Techniques for the Analysis of Urban Traffic". In: *Intelligent Transportation Systems, IEEE Transactions on* 12.3 (2011), pp. 920–939.

[Fel10] P. FELZENSZWALB, R. GIRSHICK, D. MCALLESTER, and D. RAMANAN: "Object Detection with Discriminatively Trained Part-Based Models". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 32.9 (2010), pp. 1627–1645.

[For12] D. A. FORSYTH and J. PONCE: *Computer Vision: A Modern Approach*. 2nd ed. Pearson Education Limited, 2012. ISBN: 0273764144.

[Haa99] M. HAAG and H.-H. NAGEL: "Combination of Edge Element and Optical Flow Estimates for 3D-Model-Based Vehicle Tracking in Traffic Image Sequences". In: *International Journal of Computer Vision* 35.3 (1999), pp. 295–319.

[Hod12] M. HODLMOSER, B. MICUSIK, M.-Y. LIU, M. POLLEFEYS, and M. KAMPEL: "Classification and Pose Estimation of Vehicles in Videos by 3D Modeling within Discrete-Continuous Optimization". In: *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2012*. 2012, pp. 198–205.

[Kim13] S. KIM, J. SHI, A. ALFARRARJEH, D. XU, Y. TAN, and C. SHAHABI: "Real-Time Traffic Video Analysis Using Intel Viewmont Coprocessor". In: *Databases in Networked Information Systems*. Ed. by A. MADAAN, S. KIKUCHI, and S. BHALLA. Vol. 7813. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, pp. 150–160. ISBN: 978-3-642-37133-2. DOI: 10.1007/978-3-642-37134-9_12. URL: http://dx.doi.org/10.1007/978-3-642-37134-9_12.

[Kol93] D. KOLLER: "Moving Object Recognition and Classification based on Recursive Shape Parameter Estimation". In: *In Israel Conference on Artificial Intelligence, Computer Vision*. 1993, pp. 27–28.

[Lan12] J. H. LAN, M. GUO, and X. J. LIU: "Video Event Detection Technology and its Application in Intelligent Transportation". In: *Applied Mechanics and Materials* 198-199 (2012), pp. 1225–1230.

[Pat07] I. PATRAS, E. A. HENDRIKS, and R. L. LAGENDIJK: "Probabilistic confidence measures for block matching motion estimation". In: *Circuits and Systems for Video Technology, IEEE Transactions on* 17.8 (2007), pp. 988–995.

[Zha00] Z. ZHANG: "A flexible new technique for camera calibration". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22.11 (2000), pp. 1330–1334.

[Zha10] Z. ZHANG, K. HUANG, T. TAN, and Y. WANG: "3D Model Based Vehicle Tracking Using Gradient Based Fitness Evaluation under Particle Filter Framework". In: *Pattern Recognition (ICPR), 2010*. 2010, pp. 1771–1774.

Corresponding author: Adrian Fazekas, RWTH Aachen University, Institute for Highway Engineering, 52074 Aachen, Germany, phone: +49 241 8025227, fazekas@isac.rwth-aachen.de

Exploiting Vehicle Communication with Infrastructures for Accurate Positioning

Gennaro Nicola Bifulco, Francesco Galante, Luigi Pariota, Spena Maria Russo
University of Naples Federico II

Abstract

In this paper we propose a modelling framework and an associated algorithm for the positioning of vehicles by means of a low-cost 802.15.4 wireless sensor network (WSN). The performances of the proposed approach are simulated with respect to different positioning algorithms and different steps and schemes for sensor distribution. The role of signal strength noise is also made explicit. Several simulations are run and the obtained results are shown and discussed.

Keywords: ITS, ADAS, Positioning, Wireless Sensor Networks, IEEE 802.15.4

1 Introduction

This work aims to estimate the suitability of dynamic positioning of vehicles by exploiting a road-side low-cost Wireless Sensor Network (WSN). Positioning enables Intelligent Transportation Systems (ITS) applications, as navigation services (ATIS – Advanced Traveller Assistance Systems) or Advanced Driving Assistance Systems (ADAS). Positioning can be obtained by two main techniques [Aon98], using absolute sensors such as GPS (Global Positioning Systems), and dead-reckoning ones (gyroscopes, accelerometers, etc.). In dead-reckoning the positioning error grows over time as a result of successive small uncertainties due to the integral error. GPSs work well with no overhead obstructions, in absence of multipath propagation phenomena, a sufficient number of satellites in the line of sight and a favourable Geometrical Dilution Of the Precision (GDOP); measures are quite accurate in many conditions and stable in the long term; in these conditions (and not relying on differential correction) the error of GPS is claimed to be in the magnitude of 10 metres [Abb99]. Otherwise, the GPS measurements can be lost (e.g.: tunnels) or dramatically inaccurate (e.g. canyons).

In this paper we analyse a third positioning technique via a low-cost 802.15.4 WSN based on a ZigBee specification. WSNs are widely used in many applications ([Yic08]). There is a large literature both for 2D and 3D positioning, mainly for indoor environments where GPS is problematic or impossible. Positioning from a WSN can be inferred from the Time of Arrival

(ToA) of the transmission, the Time Difference of Arrival (TdoA) or the Received Signal Strength Indicator (RSSI). In this paper we adopt the RSSI technique in an outdoor environment. We focus on positioning accuracies that can be obtained depending on different distribution schemes of the sensors and different positioning algorithms. The approach of this work is numerical, in the sense that a WSN is supposed to be in place (with different possible configurations) and experimental simulations are carried out.

2 The positioning model and the proposed algorithms

The WSN adopt the 802.15.4 standard with the ZigBee specification. In particular, without loss of generality, we refer to the CC2530 chipset by Texas Instruments. The nodes are placed road-side with a known pattern; given their role in positioning, they are called *anchor nodes* and are uniquely identified by a known MAC (Media Access Control) Address and an associated known position. A vehicle cruises in such an environment; it is equipped with a receiving beacon and knows MAC addresses and positions of anchor-nodes. Only the *polling* mechanism between nodes is exploited, and it is not required that a communication is actually established. Indeed an estimate of the distance can be made just using the RSSI, on the base of an equation given by Texas Instruments and adopted in several experiments and analyses (e.g.: [Zha09] and [Qia12]), :

$$RSSI = g(d) = -(10nlog_{10}d - A) \tag{1.a}$$

$$d = \gamma(RSSI) = 10^{\left(\frac{A-RSSI}{10\ n}\right)} \tag{1.b}$$

where n is a propagation constant, d is the transmission distance in metres, A is the received strength when the transmission distance is 1 metre, g(·) is the function of RSSI with respect to distance, γ(·) is the inverse function. Equation (1.a) is graphically shown in Figure 1.

Parameters in equation (1.a) are empirically estimated; moreover, many biases disturb the signal, including multipath phenomena. Thus the measured RSSI and the related estimated distance are random variables:

$$RSSI^M = RSSI + \varepsilon \quad \Rightarrow \quad d^M = d + \eta \tag{2}$$

where E[ε] = 0, E[$RSSI^M$] = RSSI, Var[$RSSI^M$] = Var[ε] = σ^2_ε and Var[d^M] = Var[η] = σ^2_η. Note also that, given the standard deviation of ε, that of η can be approximated as:

$$\sigma_\eta \cong \sigma_\epsilon \cdot \left[\frac{d\gamma(\cdot)}{d\ RSSI}\right]_{E[RSSI^M]} = \sigma_\epsilon \left[10^{\left(\frac{A-RSSI}{10\ n}\right)} \cdot ln(10)\right] \tag{3}$$

Here we will assume that ε is normally distributed with zero mean and known (fixed and independent on distance) standard deviation σ_ε. This means that, according to equation (3), the standard deviation σ_η depends on the signal strength measured on average at the considered position, as shown by Figure 2 below. Note that σ_η rapidly grows toward extreme values as the signal strength decreases.

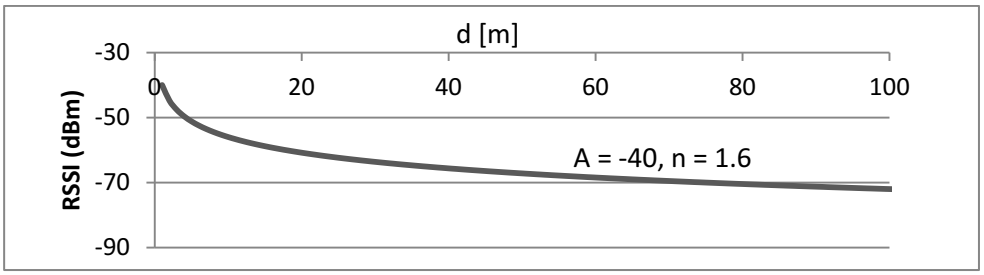


Figure 1: RSSI as a function of the transmission distance (ideal case)

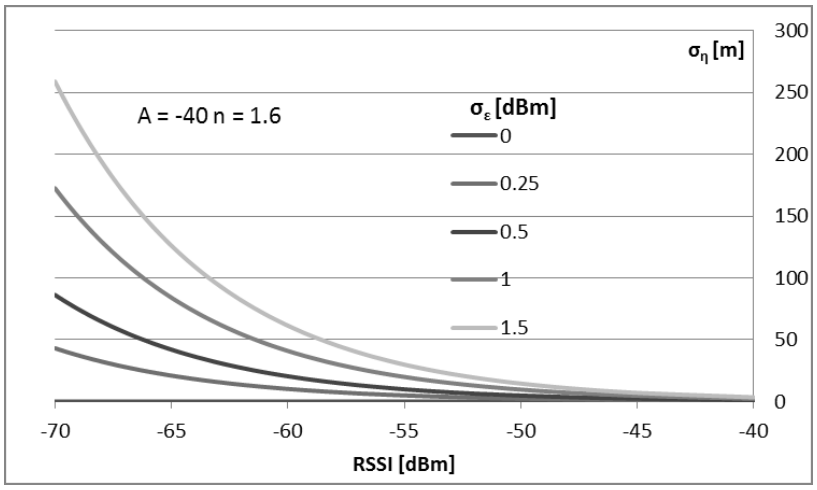


Figure 2: standard deviation of the estimated distance as a function of the measured signal, for different variances of signal strength dispersion.

Several algorithms can be used for positioning based on *anchor nodes*. Obviously, due to equation (2), a random error is introduced when the measured $RSSI_m$ is adopted in equation (1.b). As a result, error-minimising algorithms are required. Among the existing algorithms, MinMax and Maximum Likelihood are considered here. The MinMax algorithm is as introduced in [Sri02], and is a heuristic approach. The main idea is to construct a bounding box for each detected anchor node where the anchor node is the centre of the bounding-box, and its side is equal to twice the estimated distance. Thus the bounding box of an anchor node i is created by adding and subtracting the estimated distance, d^M_i, from the anchor position (x_i, y_i):

$$\left[x_i - d_i^M, y_i - d_i^M\right] \times \left[x_i + d_i^M, y_i + d_i^M\right] \quad (4)$$

The estimated position is computed as the centre of the intersection of all the bounding boxes, obtained by means of the maximum of all minima of coordinates and the minimum of all maxima of coordinates [Lan03]:

$$\left[\max(x_i - d_i^M), \max(y_i - d_i^M)\right] \times \left[\min(x_i + d_i^M), \min(y_i + d_i^M)\right] \quad (5)$$

As an alternative, the Maximum Likelihood method can be used. It is known to be equivalent to the Least Squares (LS) problem obtained by minimising the sum of squares of all differences between measured and estimated distances, where the estimated distances depend on the known positions of anchor points (x_i, y_i) and the unknown position (x_0, y_0) of

the vehicle. These differences (for any of the detected anchor points) can be written as:

$$f_i(x_0, y_0) = \left[d_i^M - \sqrt{(x_i - x_0)^2 + (y_i - y_0)^2} \right] \quad (6)$$

Equation (6) above assumes that all discrepancies from measured distances ($d^{M}_{i}=\gamma(RSSI^{M}_{i})$) and unknown distances have the same variance; equation (3) and Figure 2 show that such an assumption is really too rough, and the variance σ_η should be estimated at any given point. Thus, a Generalised Least Squares (GLS) algorithm should be implemented:

$$f_i(x_0, y_0) = \frac{d_i^M - \sqrt{(x_i - x_0)^2 + (y_i - y_0)^2}}{\sigma_\eta\ (RSSI_i)} \quad (7)$$

However, the GLS algorithm should rely on the mean $RSSI_i$ value in order to compute the variance σ_η or, at least, on a good estimate of it. If the vehicle is stationary, the estimate could be made by averaging several measures $RSSI^{M}_{i,j}$, where *i* represents a given vehicle position and *j* represents a trial of the sample of measures.

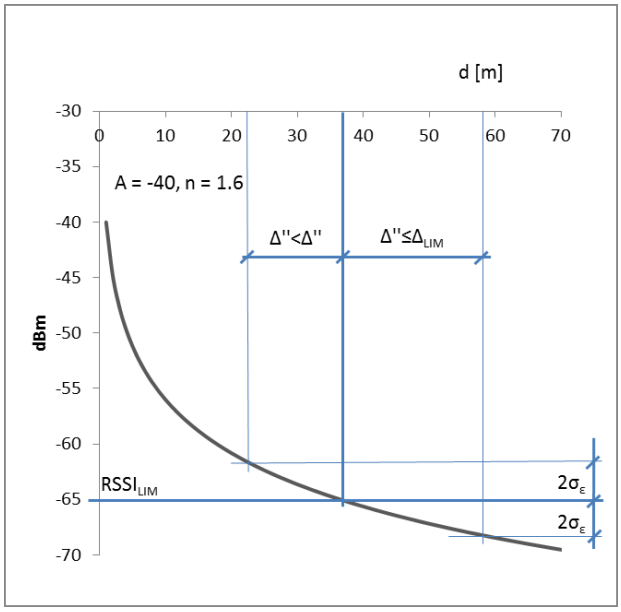


Figure 3: RSSI threshold for limiting the effect of signal strength noise.

Unfortunately, given that the vehicle cruises, the sampling rate should be really high in order to consider the trial *j* as representative of the same position *i*; in practice, this is unlikely to occur. Thus, we forgo applying the GLS and opt for the LS algorithm in an enhanced version that tries to limit the error induced by excessively feeble signals. This is not done by explicitly considering the associated variance, but by *cutting-off* from the computation an excessively feeble anchor point. This mimics the behaviour of the GLS where, consistent with equation 7, the measures with extremely low (and unreliable) RSSI give a negligible contribution. In practice, in order to control the influence of unreliable distance estimates affected by a potentially *disruptive noise*, we limit the accepted signal strength to a value ($RRSI_{LIM}$) such that the noise in the measured signal strength results in a maximum distance error in 95% of cases (Δ'', see Figure 3) no greater than a given value Δ_{LIM}, that can be heuristically fixed as the minimum distance between any pair of anchor points. In

conclusion, we propose to deal with the model presented in this section by means of two different algorithms and, for each, two different variants, as shown in Table 1.

Table 1: Algorithms used in our simulation.

Error minimisation approach	**Cut-off of disruptive noise**	
	NO	**YES**
MinMax	Algorithm 1	Algorithm 3
Least Square	Algorithm 2	Algorithm 4

3 The simulated scenarios

For the purposes of this paper we considered a straight 1-km section of a typical Italian motorway consisting of two carriageways, each with three lanes 3.75 m wide (*nl*=3), an emergency lane 3.0 m wide, and a central reservation 3.4 m wide. As a result, the total motorway width is 31.9 m. A vehicle cruises with uniform speed on a given lane (the emergency lane is not considered, without loss of generality). Sensors along the motorway were placed according to three different schemes (*ns*=3) for which an offprint is shown in Figure 4 below. It is worth noting that the linear density of the sensors is the same for all schemes (one sensor every *p* metres, where *p* here stands for the *step* of the sensor distribution). Three different steps (*np*=3, with *p*=10, 20 and 30 metres) were tested in our work in order to understand how they impact on positioning accuracy. In the simulation, as the vehicle cruises, the distance from any anchor node is known. Thus the RSSI that should be detected by the vehicle is known in accordance with equation (6.a). In order to simulate the signal strength bias, the RSSI is perturbed with a random draw of ε, as in equation (2). Different values were tested for the signal strength bias σ_ε, in different simulations; they are in number of *ne*=*5* and their values are 0 (no randomness), 0.25, 0.50, 1.0 and 1.5 dBm. It is worth noting that the higher of the standard deviations also implicitly (and roughly) accounts for serious multipath phenomena. The perturbed RSSI is assumed to be measured by the vehicle and used in equation (1.b). As in the real world, our vehicle is not able to detect all signals; anchor points for which the signal is less than -70 dB (that correspond to an estimated distance greater than 75 m) are considered as not being detected or as neglected by the vehicle. Moreover, a further threshold could be considered to implement the variant oriented to cut-off anchor points with potential disruptive noise, as discussed in the previous section with reference to figure 3. The value of Δ_{LIM}, the minimum distance between any pair of anchor points, depends on both the scheme and the sensor distribution step and is computed by geometrical considerations.

Generating all possible combinations leads to (*ns*×*np*×*ne* = *3*×*3*×*5*) 45 different scenarios. Moreover, in each scenario the positions of the vehicles in each of the three lanes (*nl*=3) were considered. Finally, the position was estimated by using the MinMax, and Least Square algorithms (*na*=2) with the variants in Table 1 of section 2 (*nw*=2, with and without considering disruptive noises). Estimation accuracy was evaluated differently for the two coordinates (*nc*=2) x and y. From all this it results that (45×*nl*×*na*×*nw*×*nc*=*45*×*3*×2×*2*×*2*) 1080 scenarios have to be run and analysed and for each scenario the positioning algorithm is

applied several times, each 0.1 seconds (10 Hz) of the simulation.

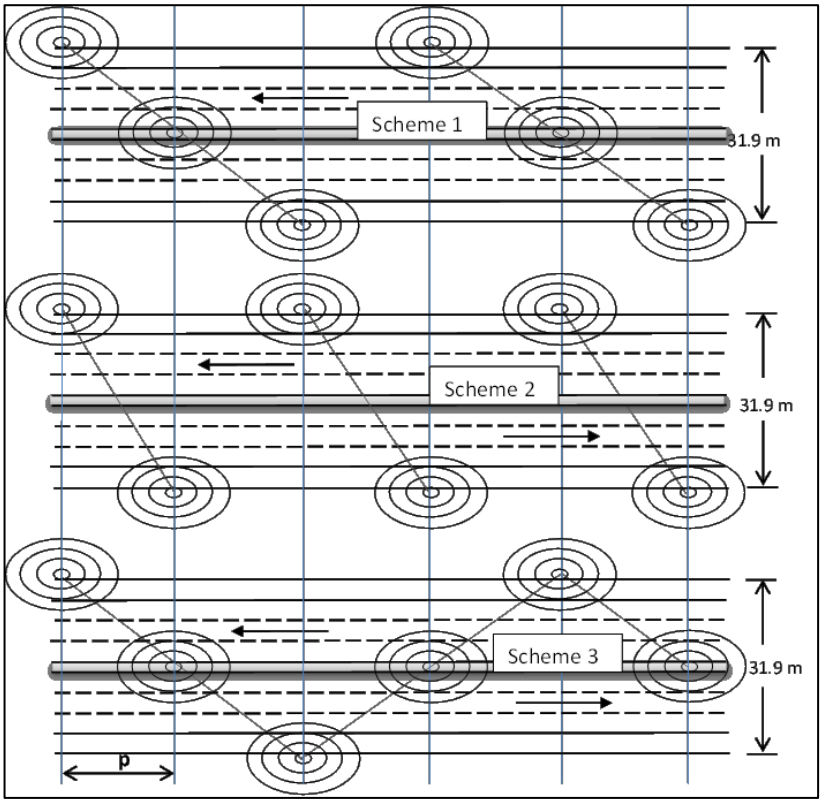


Figure 4: Schemes for sensor distribution.

4 Results of the simulations and discussion

In order to understand the more relevant variables of the process, evaluation of the positioning performance was based on the root mean square error (RMSE):

$$RMSE_X = \sqrt{\frac{\sum_{i=1}^{n}\left(x_{E,i} - x_R\right)^2}{n}} \qquad RMSE_y = \sqrt{\frac{\sum_{i=1}^{n}\left(y_{E,i} - y_R\right)^2}{n}} \tag{8}$$

where $x_{E,i}$ [$y_{E,i}$] is the estimated x [y] coordinate, x_R [y_R] is the actual x [y] coordinate of the vehicle and n is the number of estimated positions along the vehicle's trajectory. Positioning along x is defined as longitudinal positioning, while positioning along y transversal.

Another measure to be taken under control is the (horizontal) DOP (Dilution Of Precision), the HDOP. It measures the occurrence of unfavourable reciprocal position of the anchor points with respect to the vehicle. If the DOP is high it means that, given the number of anchor points considered, they are not well distributed in the space around the vehicle, and this reduces the accuracy. Of course, given two configurations both with acceptable DOP, the most convenient is the one with the higher RSSI and not with the lower DOP.

In order to attain a robust estimation of the RMSE, and considering the randomness of the signal strength bias, 100 simulations were carried out for each of the scenarios (except for those in which σ_ε was set to 0). Figure 5 shows the results of the longitudinal and the transversal positioning for the different algorithms (and variants) as described in Table 1.

Simulated scenarios are numbered in descending order of performance (increasing RMSE). Longitudinal positioning is more successful than transversal; this could be due to the total width of the roadway that is not negligible and comparable with the less frequent of the sensor distribution steps.The LS algorithm without cut-off of disruptive noise is the worst performing. The best performing algorithm, overall for longitudinal and transversal positioning, is the algorithm number 3 (LS with cut-off of disruptive noise), Algorithms 2 and 4 (MinMax with and without cut-off of disruptive noise) perform very similarly and are definitely robust with respect to the signal strength noise.

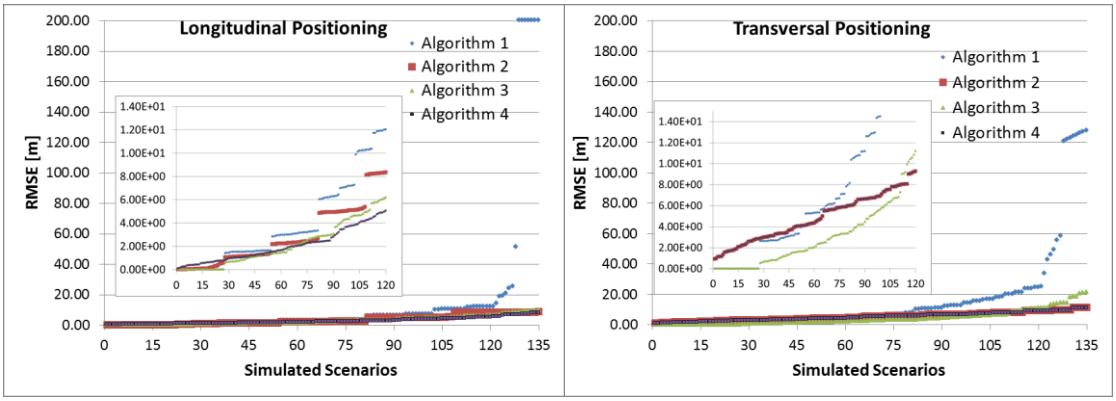


Figure 5: Performance of the algorithms tested.

The analyses that will follow will refer only to algorithm 3. Longitudinal positioning seems to be scarcely affected by the distribution schemes shown in Figure 4, while scheme 2 clearly performs best for transversal positioning (see Figure 6). Given that the accuracy of transversal positioning is more problematic than the longitudinal one, we will adopt for all further analyses scheme 2.

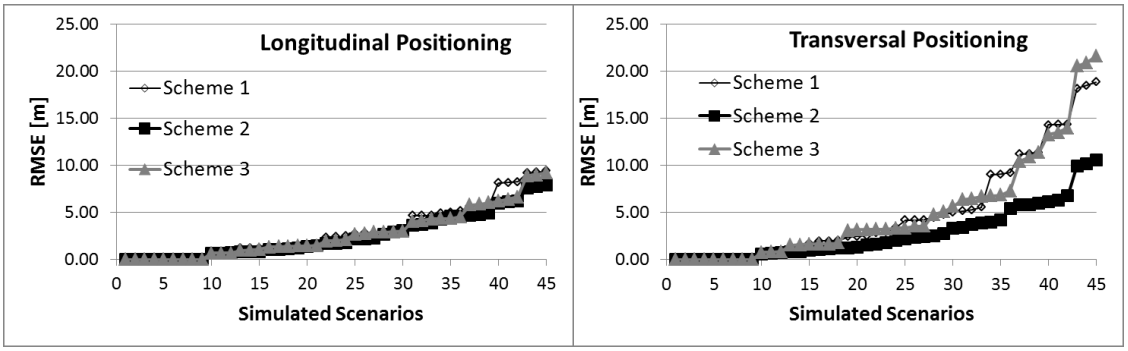


Figure 6: Performance of the distribution schemes.

Once algorithm 3 and scheme 2 have been proved to perform best, the role of the distribution step and of the signal noise can be investigated in Figure 7, where low variances of the signal strength have not been shown, as in these cases the RMSE is really low and almost independent on the step of the sensor distribution. Three RMSE points are depicted for each step and, for each standard deviation, these show the RMSE for three different lanes (the right, middle and left lanes). In the case of longitudinal positioning, assume that the system is in place with step 30 m and signal strength standard deviation σ_ε =0.5 dBm. Now, according to Figure 7, assume that the signal strength standard deviation moves toward 1 dBm; thus the RMSE moves from 2 to 4.5 m. If the distribution step now decreases to 20 m (or 10 m), the RMSE decreases from 4.5 to 4 (or 3) m, we are thus not able to compensate

what we have lost. In other words, with respect to longitudinal positioning, a controlled variance of the signal strength is more important than a more frequent distribution of the sensors, and hence a step distribution of 30 m ensures a good cost-benefit performance. Things are different for the transversal distribution. First of all, it is evident that the RMSE depends more clearly on the lane on which the vehicle cruises. This is more evident for step 10 and, in any case, the dispersion over lanes increases as the variance of the signal strength increases. In the case of transversal positioning, if one considers a distribution step of 30 m, the increase in the RMSE due to an increased variance of the signal strength is dramatic but can be quite well compensated by moving to a 20 m step distribution. The same compensation does not hold if from the 20 m step one moves to a 10 m step; moreover, the 20 m step is less sensitive than the 10 m one on the lane it refers to. In summary, for transversal positioning, the 20 m step distribution seems to be the best cost-benefit balance (and still a relatively small variance in the signal strength is a key variable for successful positioning).

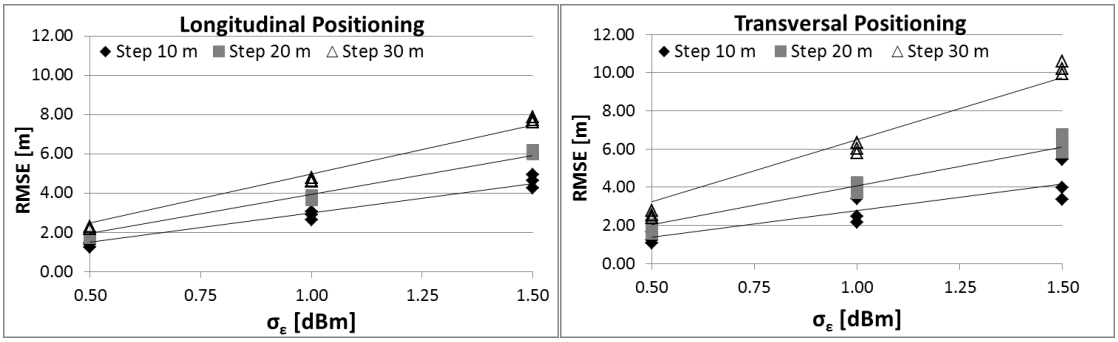


Figure 7: Role of distribution step and RSSI noise.

The employed algorithms, as well as the relative performances in terms of longitudinal and transversal positioning, have been analysed in terms of DOP too. Figure 8 shows that the DOP computed for the algorithms with and without the cutting-off the extremely low RSSI is in both cases excellent (less than 3.5), even if slightly better in case the cutting-off is adopted. With respect to the algorithm with cutting-off, Figure 9 shows that the values of DOP are always excellent, that the transversal DOP is slightly sensitive to the lane and that the transversal positioning is slightly better in terms of DOP than the longitudinal one. This last consideration is contradictory with the better accuracy of longitudinal positioning that has been evidenced by Figure 7. Of course, as already said, if the value of DOP is good, then the accuracy depends on the RSSI and is constant with respect to the DOP.

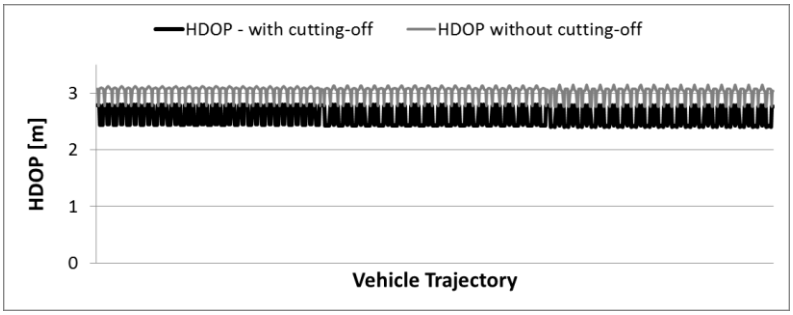


Figure 8: DOP evaluated with and without anchor-nodes cutting-off.

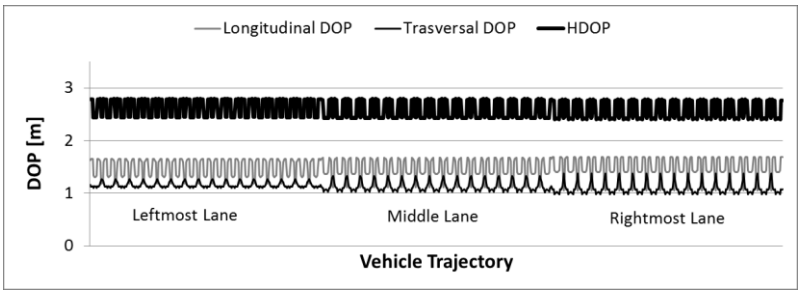


Figure 9: Longitudinal, transversal DOP and H (total) DOP.

5 Conclusions

We tested the use for vehicle positioning of a low-cost and low-consumption 802.15.4 WSN, based on the ZigBee specification. The approach allows for positioning vehicles with sufficient accuracy, using appropriate algorithms, distribution schemes and steps. A Least Squares algorithm with appropriate cut-off of potentially disruptive sensors (because of high distance and feeble signal) gives good results, without need to use a generalised least squares method. The positioning accuracies are comparable with those of GPS, and the system can be used where the GPS cannot (canyons, tunnels, etc.). Some results obtained for indoor environments [Gol10] are also confirmed for the outdoor one; a cost-effective step for the distribution of the sensors is 20 m. Such a step is able to minimise disturbance induced by a noisy received signal strength intensity. Interestingly, the solution proposed entails a low power consumption and a low cost. Assuming that the investment cost is 30 euro per anchor point, the selected distribution step ensures that an infrastructure can be equipped at a cost of 1500 euro/km, which is a reasonable investment cost.

6 Acknowledgments

This research was partly carried out within the Italian PON (Programma Operativo Nazionale) 2007-13 - Research Project B61H110004000005 *DRIVE IN²*. Preliminary tests were carried out by using the software PreScan (Simulation of ADAS and Active Safety) by Tass International (http://www.tassinternational.com/prescan). We would also like to thank Vittorio Marzano for fruitful discussions. Finally, many thanks to the anonymous reviewers for their contribution to the improvement of the quality of the paper.

References

[Abb99] H. ABBOTT and D. POWELL: "Land-vehicle navigation using GPS". In: *Proceedings of the IEEE* 87.1 (1999), pp.145–162. DOI: 10.1109/5.736347.

[Aon98] T. AONO, K. FUJII, S. HATSUMOTO, and T. KAMIJIA: "Positioning of vehicle on undulating ground using GPS and dead reckoning". In: *IEEE International Conference on Robotics and Automation*. Vol. 4. 1998, pp.3443–3448.

[Gol10] E. GOLDONI, A. SAVIOLI, M. RISI, and P. GAMBA: "Experimental analysis of RSSI-based indoor localization with IEEE 802.15.4". In: *2010 European Wireless Conference (EW)*. Lucca, Italy, Apr. 12–14, 2010, pp. 71–77.

[Lan03] K. LANGENDOEN and N. REIJERS: "Distributed localization in wireless sensor networks: a quantitative comparison". In: *Series Computer Networks* 43 (2003), pp. 499–518.

[Qia12] Q. DONG and W. DARGIE: "Evaluation of the reliability of RSSI for indoor localization" In: *International Conference Wireless Communications in Unusual and Confined Areas (ICWCUCA)*. Aug. 28–30, 2012, pp. 1–6. DOI: 10.1109/ICWCUCA.2012.6402492.

[Sri02] M. SRIVASTAVA, A. SAVVIDES, and H. PARK: "The bits and flops of the n-hop multilateration primitive for node localization problem". In: *IEEE Sensors Journal* 7 (Sept. 2002), pp. 557–561.

[Yic08] J. YICK, B. MUKHERJEE, and D. GHOSAL: "Wireless sensor network survey". In: *Series Computer Networks* 52 (2008), pp. 2292–2330.

[Zha09] ZHANG JIANWU and ZHANG LU: "Research on distance measurement based on RSSI of ZigBee". In: *International Colloquium Computing, Communication, Control, and Management (ISECS)*. Vol. 3. Aug. 8–9, (2009), pp. 210–212. DOI: 10.1109/CCCM.2009.5267883.

Corresponding author: Gennaro Nicola Bifulco, University of Naples Federico II, Department of Civil and Environmental Engineering, 80125 Napoli, Italy, phone: +39 081 76 83883, e-mail: gennaro.bifulco@unina.it.

Location Forwarding for Dense Urban Environments

Alireza Ghods, Stefano Severi, Giuseppe Abreu
Jacobs University

Abstract

This paper addresses the problem of vehicle position estimation in dense urban environments, where traditional Global Positioning System (GPS)-based localisation techniques are severely affected by non line-of-sight (NLOS) signal propagation and multipaths presence. Assuming that GPS signals are fairly received only by a very small fraction of vehicles at the border of the urban environment, we propose a solution based on vehicle to vehicle (V2V) communication to propagate this information to the whole network. As a consequence, this multihop scheme allows vehicles to estimate its own position collecting their cumulative distances to border vehicles.

Finally we introduce a new analytical framework to verify the fundamental performance of the proposed solution in term of position estimate error bounds, jointly considering the uncertainty introduced by the multihop process and by the GPS localization.

Keywords:

1 Introduction

In this paper we consider a typical dense urban environment in which global positioning system global positioning system (GPS) signal is weakened by buildings and other similar obstacles that cause non line-of-sight non line-of-sight (NLOS) and multipath propagation. Under these circumstances, vehicle localization precision can be severely affected, resulting in inaccurate position estimations.

We therefore propose a new method to augment the localization information, and consequently to increase the reliability of the process, introducing a vehicle to vehicle (V2V) multihop communication scheme, that does not rely on any fixed infrastructure. Vehicle that are at the border of the dense urban environment are assumed to have a stronger GPS signal with respect to those inside, hence they have a more accurate position estimate. Taking advantage from this, we will use these vehicles as anchor nodes for a multihop scheme

to improve the localization accuracy of any possible vehicle driving inside the dense urban environment.

Two main problem are therefore considered and solved: the lack of direct path between target and anchors and the uncertainty of anchor position estimates, which is then inevitably propagated during the localization process. We combine, as a consequence, a multihop localization scheme [Sev12] with a unified framework for target localization with anchors uncertainty.

We finally compute the fundamental limits of this combined approach via newly reformulated Fisher information matrix (FIM) that allow us to consider jointly the uncertainty introduced by the multihop process and by the GPS localization, and we validate the theoretical model via simulations[1].

2 Cooperative Network Localization

2.1 Cramèr-Rao lower bound of Target Specific Approach

A *vehicular network* is understood as a set of N nodes (vehicles), whose η-dimensional coordinate (column) vectors are indexed as $[\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{n_\mathrm{T}}, \boldsymbol{a}_{n_{\mathrm{T}+1}}, \ldots, \boldsymbol{a}_N]$. In particular, for bidimensional case, the element of the i-th vehicle in the dimension "x" and "y" are denoted by x_i and y_i, respectively. While the position of the first n_T vehicles (referred to as *targets*) is unknown, the location of a small fraction $(K = N - n_\mathrm{T})$ of vehicles, (*anchors*), out of the dense urban area, can be determined via GPS. The uncertainty of the position estimate of the k-th anchor $\boldsymbol{a}_k$ is described by the η-by-η covariance matrix $\boldsymbol{\Sigma}_k$.

Let d_{ij} denote the mutual distance between nodes i and j, such that, $d_{ij} \triangleq \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_j\| = \sqrt{\langle \boldsymbol{\theta}_i - \boldsymbol{\theta}_j, \boldsymbol{\theta}_i - \boldsymbol{\theta}_j \rangle}$, where $\|\cdot\|$ and $\langle\cdot\rangle$ denote Euclidean norm and inner product, respectively. We refer to the *neighborhood set* of the i-th node $\mathcal{N}_i$ as the group of nodes j in the vicinity of i, to which distance measurements[2] $\tilde{d}_{ij}$ of d_{ij} can be obtained directly. For further convenience, define also the *neighborhood function* $I_{\mathcal{N}_i}(j)$, which takes value 1 if the j-th node belongs to $\mathcal{N}_i$ and 0 otherwise.

Having defined this notation, we want now to independently localize each single target inside the dense urban environment, only relying on the position information sent by anchor vehicles and propagated by other targets within the network. To clarify, let a route between a generic target at $\boldsymbol{\theta}$ (we drop the subscript i without any loss of generality) and an anchor at location $\boldsymbol{a}_k$ involves n_k hops. We assume a distance measurement $\tilde{d}_k$, related to the k-th hop within the route, being affected by zero-mean Gaussian error (see [Jou08] and [Nic09] for a general description of the error model) with variance σ_k^2 linearly proportional to the inverse of the signal-to-noise ratio (SNR) over the radio link between two intermediate vehicles, that is

$$\sigma_{ij}^2 \triangleq \sigma_0^2 \cdot \left(\frac{d_{ij}}{d_0}\right)^{\alpha}, \tag{1}$$

[1] We will presented simulation results in the camera-ready version of the paper.

[2] We will distinguish between *direct measurements* and *multihop distance estimates* by adding accents ˜ and ¯, respectively, to the letter denoting the referred quantity.

where $\alpha \geq 0$ and $\sigma_0^2 = 0.2/(\sqrt{10^{SNR/10}})$ is the ranging variance at a reference distance d_0. Then the measured multihop distance to the k-th anchor vehicle is

$$\bar{d}_k \triangleq \sum^{n_k} \tilde{d}_k, \tag{2}$$

where $\tilde{d}_k$'s are zero mean Gaussian random variables with variances σ_k^2, such that the variance of $\bar{d}_{ij}$ becomes

$$\bar{\sigma}_k^2 \triangleq \sum^{n_k} \sigma_k^2. \tag{3}$$

Let now $\hat{\boldsymbol{\theta}}$ denote the estimate of the location of a generic single target, obtained with basis on the set of multihop distance estimates between itself and the anchors, that is, $\bar{\boldsymbol{d}} \triangleq [\bar{d}_{n_{\mathrm{T}+1}}, \cdots, \bar{d}_N]$. Associated with $\hat{\boldsymbol{\theta}}$ there is the η-by-η covariance matrix:

$$\boldsymbol{\Omega}_{\boldsymbol{\theta}} \triangleq \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^{\mathrm{T}}\right]. \tag{4}$$

The Cramèr-Rao lower bound (CRLB)

$$\boldsymbol{\Omega}_{\boldsymbol{\theta}} \succeq \mathbf{F}_{\boldsymbol{\theta}}^{-1}, \tag{5}$$

relates the covariance matrices $\boldsymbol{\Omega}_{\boldsymbol{\theta}_i}$ to (FIM). Clearly under this classic formulation, $\mathbf{F}_{\boldsymbol{\theta}}$ depends only on the distances estimates between the target vehicle and the anchors, collected via multihop paths from the target to the anchors. Consequently the (CLRB) on the errors of $\hat{\boldsymbol{\theta}}$ ultimately depends on the error processes affecting the multihop distance estimates between nodes composing the route from node i to the anchors, without considering the uncertainty on anchor positions due to GPS localization.

We therefore augment $\boldsymbol{\theta}$ with the anchor positions, obtaining a new parameter vector

$$\boldsymbol{\Theta} = \left[\boldsymbol{\theta}^{\mathrm{T}}, \mathbf{a}_1^{\mathrm{T}}, \mathbf{a}_2^{\mathrm{T}}, \cdots, \boldsymbol{a}_K^{\mathrm{T}}\right]^{\mathrm{T}}, \tag{6}$$

that allows us to define a new $\eta(K+1)$-by-$\eta(K+1)$ covariance matrix

$$\boldsymbol{\Omega}_{\boldsymbol{\Theta}} \triangleq \mathbb{E}\left[(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})^{\mathrm{T}}\right], \tag{7}$$

where quantities denoted by $\hat{}$ denote estimates. The joint anchors-target (CLRB) now relates eq. (7) to a new (FIM) as follows:

$$\boldsymbol{\Omega}_{\boldsymbol{\Theta}} \succeq \mathbf{F}_{\boldsymbol{\Theta}}^{-1}. \tag{8}$$

2.2 Reformulated Fisher information matrix Considering Anchor Uncertainty

The (FIM) in eq. (8) can be accurately approximated by the sum of two matrices, namely $\mathbf{F}_{\mathrm{M}}$ and $\mathbf{F}_{\Sigma}$ the former accounting for uncertainty due to *multihop* measurements, and the latter

corresponding to the GPS error on anchor position estimates. We can then write:

$$\mathbf{F}_{\boldsymbol{\Theta}} \approx \mathbf{F}_{\mathrm{M}} + \mathbf{F}_{\Sigma}, \tag{9}$$

where the approximation holds whenever the anchor-to-target distances are much greater then GPS error is , i.e. when $\|\boldsymbol{\theta} - \boldsymbol{a}_k\| \gg \mathrm{tr}(\boldsymbol{\Sigma}_k), \forall\, k$.

The standard approach to compute the (FIM) is based on the derivation of the log likelihood function of the measurements [Kay93]. However, computing $\mathbf{F}_{\boldsymbol{\Theta}}$ element by element can be a complex task for large number of targets; therefore we have decided to compute $\mathbf{F}_{\mathrm{M}}$ and $\mathbf{F}_{\Sigma}$ separately, employing also a more systematically approach resulting from a simpler and faster algorithm.

In particular the (FIM) related to the uncertainty over a single target $\boldsymbol{\theta}$ equivalent to the inverse of the right term in eq. (5), can be expressed as:

$$\mathbf{F}_{\boldsymbol{\theta}} = \sum_{k \in K} \mathbf{u}_k \mathbf{u}_k^{\mathrm{T}}, \tag{10}$$

where k is the index number of the anchor and $\mathbf{u}_k$ is defined as:

$$\mathbf{u}_k = \frac{\partial \|\mathbf{a}_k - \boldsymbol{\theta}\|}{\partial \boldsymbol{\theta}} \sqrt{F_k} = \frac{1}{\bar{d}_k} \left[(x_{a_k} - x_\theta), (y_{a_k} - y_\theta)\right]^{\mathrm{T}} \sqrt{F_k}, \tag{11}$$

where

$$F_k = \frac{1}{\bar{\sigma}_k^2} \left(1 + \frac{\alpha^2\, \sigma_0^2}{2\, d_0^\alpha} (\|\boldsymbol{a}_k - \boldsymbol{\theta}\|)^{\alpha - 2} \right), \tag{12}$$

and α is the pathloss exponent, typically around 1.8 for V2V communication [Pai08].

Back to eq. (9), we can now compute the $\mathbf{F}_{\mathrm{M}}$ inserting the augmented vector $\boldsymbol{\Theta}$ into eq. (11) and (12), thus obtaining:

$$\frac{\partial \|\mathbf{a}_k - \boldsymbol{\Theta}\|}{\partial \boldsymbol{\Theta}} = \frac{1}{\sqrt{F_k}} \left[\mathbf{u}_k^{\mathrm{T}},\, \mathbf{0}_{1 \times \eta \cdot (k-1)},\, -\mathbf{u}_k^{\mathrm{T}},\, \mathbf{0}_{1 \times \eta \cdot (K-k)} \right]^{\mathrm{T}}, \tag{13}$$

where $\mathbf{0}_{n \times m}$ denotes a null matrix with n rows and m columns.

Having defined the vectors

$$\mathbf{v}_k \triangleq \sqrt{F_k} \cdot \frac{\partial \|\mathbf{a}_k - \boldsymbol{\Theta}\|}{\partial \boldsymbol{\Theta}}, \tag{14}$$

we can express

$$\mathbf{F}_{\mathrm{M}} = \sum_{k=1}^{K} \mathbf{v}_k \mathbf{v}_k^{\mathrm{T}} = \begin{bmatrix} \mathbf{A} & \mathbf{B}^{\mathrm{T}} \\ \mathbf{B} & \mathbf{C} \end{bmatrix}, \tag{15}$$

where $\mathbf{A} \triangleq \sum_{k=1}^{K} \mathbf{u}_k \mathbf{u}_k^{\mathrm{T}}$, $\mathbf{B}^{\mathrm{T}} \triangleq \left[-\mathbf{u}_1 \mathbf{u}_1^{\mathrm{T}}, \cdots, -\mathbf{u}_K \mathbf{u}_K^{\mathrm{T}} \right]$ and $\mathbf{C}$ is the block diagonal matrix with blocks given by $\mathbf{u}_k \mathbf{u}_k^{\mathrm{T}}$, that is, $\mathbf{C} \triangleq \mathrm{diag}\left(\mathbf{u}_1 \mathbf{u}_1^{\mathrm{T}}, \cdots, \mathbf{u}_K \mathbf{u}_K^{\mathrm{T}} \right)$.

The GPS error related $\mathbf{F}_{\Sigma}$ matrix is defined by

$$\mathbf{F}_{\Sigma} \triangleq \begin{bmatrix} \mathbf{0}_{M\times M} & \mathbf{0}_{M\times MK} \\ \mathbf{0}_{KM\times M} & \boldsymbol{\Sigma}^{-1} \end{bmatrix} \tag{16}$$

where $\boldsymbol{\Sigma}$ is a block-diagonal matrix $\boldsymbol{\Sigma} \triangleq \operatorname{diag}(\boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\Sigma}_K)$.

An expression of the (FIM) that also considers anchor uncertainty in contrast to eq. (5), corresponding to a generic target vehicle $\boldsymbol{\theta}$, can finally be obtained by taking the η-by-η Schur complement of $\mathbf{F}_{\mathrm{M}} + \mathbf{F}_{\Sigma}$, which yields to:

$$\begin{aligned}
\mathbf{F}_{\boldsymbol{\theta}}^{*} &= \mathbf{A} - \mathbf{B}^{\mathrm{T}} \left(\boldsymbol{\Sigma}^{-1} + \mathbf{C}\right)^{-1} \mathbf{B}, \\
&= \sum_{k=1}^{K} \mathbf{u}_k \mathbf{u}_k^{\mathrm{T}} - \sum_{k=1}^{K} \mathbf{u}_k \mathbf{u}_k^{\mathrm{T}} \left(\boldsymbol{\Sigma}_k^{-1} + \mathbf{u}_k \mathbf{u}_k^{\mathrm{T}}\right)^{-1} \mathbf{u}_k \mathbf{u}_k^{\mathrm{T}}, \\
&= \sum_{k=1}^{K} \mathbf{u}_k \left(1 - \mathbf{u}_k^{\mathrm{T}} \left(\boldsymbol{\Sigma}_k^{-1} + \mathbf{u}_k \mathbf{u}_k^{\mathrm{T}}\right)^{-1} \mathbf{u}_k\right) \mathbf{u}_k^{\mathrm{T}}, \\
&= \sum_{k=1}^{K} \mathbf{u}_k \left(1 - \mathbf{u}_k^{\mathrm{T}} \left(\boldsymbol{\Sigma}_k - \frac{\boldsymbol{\Sigma}_k \mathbf{u}_k \mathbf{u}_k^{\mathrm{T}} \boldsymbol{\Sigma}_k}{1 + \mathbf{u}_k^{\mathrm{T}} \boldsymbol{\Sigma}_k \mathbf{u}_k}\right) \mathbf{u}_k\right) \mathbf{u}_k^{\mathrm{T}}, \\
&= \sum_{k=1}^{K} \mathbf{u}_k \left(1 - \mathbf{u}_k^{\mathrm{T}} \boldsymbol{\Sigma}_k \mathbf{u}_k + \frac{\mathbf{u}_k^{\mathrm{T}} \boldsymbol{\Sigma}_k \mathbf{u}_k \mathbf{u}_k^{\mathrm{T}} \boldsymbol{\Sigma}_k \mathbf{u}_k}{1 + \mathbf{u}_k^{\mathrm{T}} \boldsymbol{\Sigma}_k \mathbf{u}_k}\right) \mathbf{u}_k^{\mathrm{T}}, \\
&= \sum_{k=1}^{K} \mathbf{u}_k \left(1 - \nu_k + \frac{\nu_k^2}{1 + \nu_k}\right) \mathbf{u}_k^{\mathrm{T}}, \\
&= \sum_{k=1}^{K} \frac{1}{1 + \nu_k} \mathbf{u}_k \mathbf{u}_k^{\mathrm{T}},
\end{aligned} \tag{17}$$

where we have made use of the Sherman-Morrison formula and implicitly defined $\nu_k \triangleq \mathbf{u}_k^{\mathrm{T}} \boldsymbol{\Sigma}_k \mathbf{u}_k$.

Therefore the target (FIM) considering the anchor uncertainty will be

$$\mathbf{F}_{\boldsymbol{\theta}}^{*} = \sum_{k=1}^{K} \frac{\mathbf{u}_k \mathbf{u}_k^{\mathrm{T}}}{1 + \nu_k}. \tag{18}$$

3 Error Bounds

From eq. (18) and eq. (5) we can employ the (CLRB) to lower bound the new covariance matrix

$$\boldsymbol{\Omega}_{\boldsymbol{\theta}}^{*} \succeq \mathbf{F}_{\boldsymbol{\theta}}^{*-1}, \tag{19}$$

where, for a bidimensional case, we have

$$\boldsymbol{\Omega}_{\boldsymbol{\theta}}^{*} = \begin{bmatrix} \sigma_{\mathrm{x}}^2 & \sigma_{\mathrm{xy}} \\ \sigma_{\mathrm{xy}} & \sigma_{\mathrm{y}}^2 \end{bmatrix}. \tag{20}$$

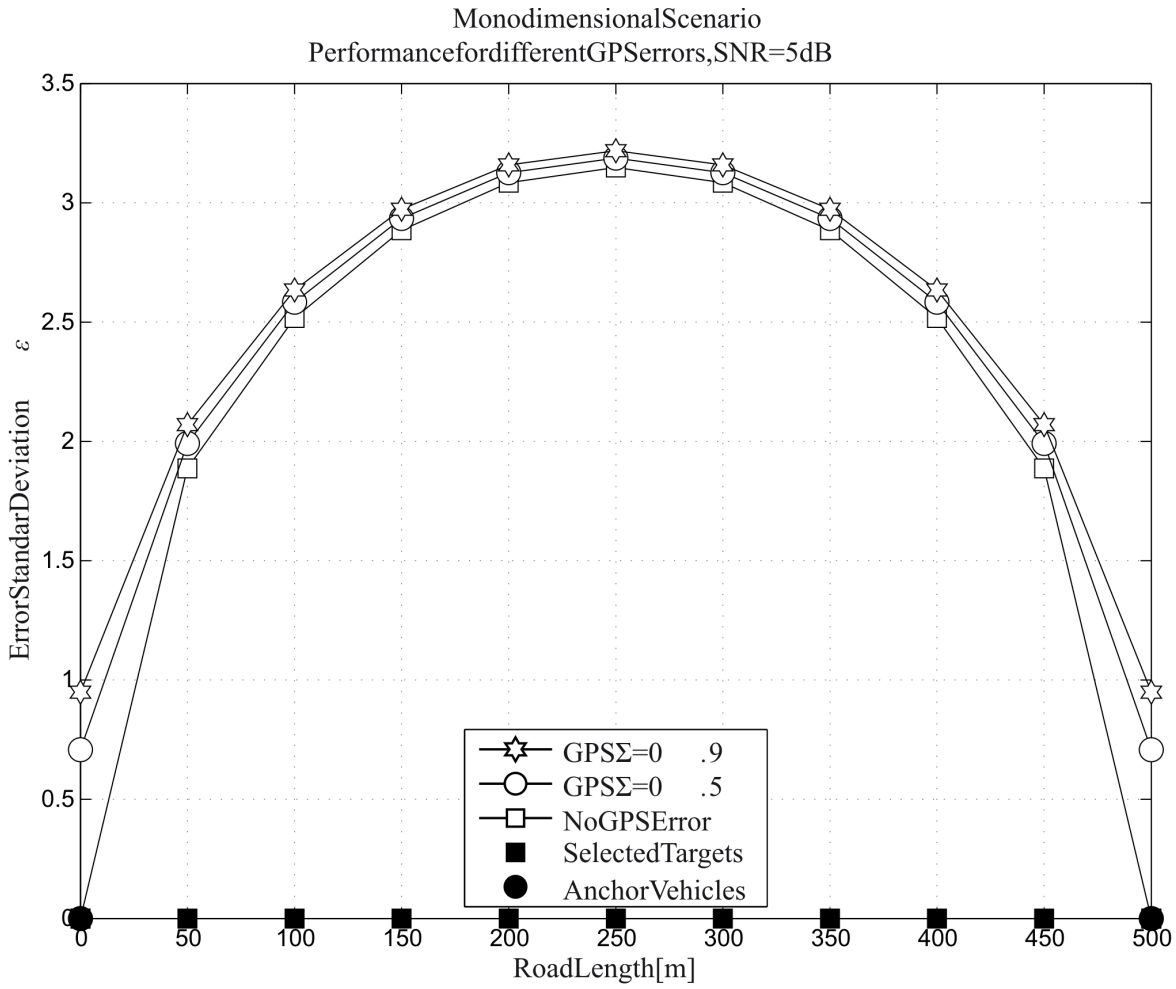


Figure 1: CRLB for Accurate vs Inaccurate Anchor Estimation Given Fixed SNR

It is well known that the directions of maximum *dispersion* in the space for the random vector $\hat{\boldsymbol{\theta}}$ are proportional, up to a factor κ, to the eigenvalues associated to $\boldsymbol{\Omega}_{\theta}^{*}$.

Specifically, the axis of the ellipse that better describes such a dispersion in the space are given by $2\sqrt{\kappa\lambda_1}$, $2\sqrt{\kappa\lambda_2}$ respectively [Sev12], where

$$\lambda_1 \triangleq \frac{1}{2}\left[\sigma_{\mathrm{x}}^2 + \sigma_{\mathrm{y}}^2 + \sqrt{(\sigma_{\mathrm{x}}^2 - \sigma_{\mathrm{y}}^2)^2 + 4\sigma_{\mathrm{xy}}^2}\right], \tag{21}$$

$$\lambda_2 \triangleq \frac{1}{2}\left[\sigma_{\mathrm{x}}^2 + \sigma_{\mathrm{y}}^2 - \sqrt{(\sigma_{\mathrm{x}}^2 - \sigma_{\mathrm{y}}^2)^2 + 4\sigma_{\mathrm{xy}}^2}\right]. \tag{22}$$

In presence of Gaussian random vectors, the proportionality factor κ is related to probability P_e that the target $\hat{\theta}$ is enclosed in ellipse, and is given by

$$\kappa = -2\ln(1 - P_e). \tag{23}$$

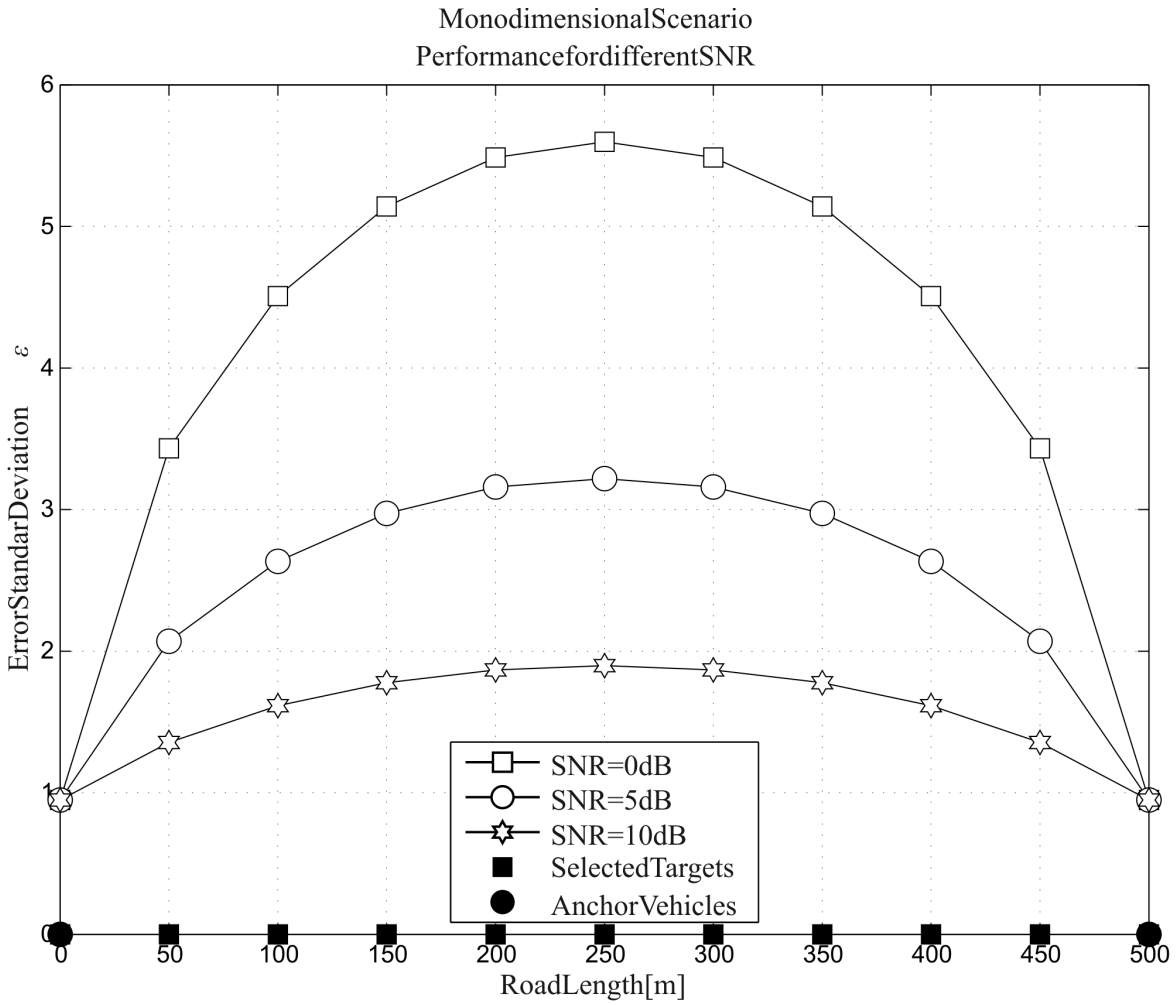


Figure 2: CRLB for Varying SNR Given Anchor Position Uncertainty

The Fisher Ellipse for the generic target $\hat{\boldsymbol{\theta}}$ is described by the following equation [Tor84]

$$\begin{aligned} & \frac{[(x-\theta_{\mathrm{x}})\cos\gamma+(y-\theta_{\mathrm{y}})\sin\gamma]^2}{\kappa\cdot\lambda_1} \\ + \quad & \frac{[(x-\theta_{\mathrm{x}})\sin\gamma-(y-\theta_{\mathrm{y}})\cos\gamma]^2}{\kappa\cdot\lambda_2}=1, \end{aligned} \tag{24}$$

where γ is the *rotation angle* used to describe the offset between the principal axis for the ellipse and reference axis of the system, and it is given by

$$\gamma \triangleq \frac{1}{2}\arctan\left(\frac{2\sigma_{\mathrm{xy}}}{\sigma_{\mathrm{x}}^2-\sigma_{\mathrm{y}}^2}\right). \tag{25}$$

4 Results and Comparison

In this section we evaluate the theoretical performance of the proposed multihop solution for dense urban scenario. We have therefore considered a road of a total length of 500 meters, which is supposed to go through a dense environment where GPS signal is either poor or not

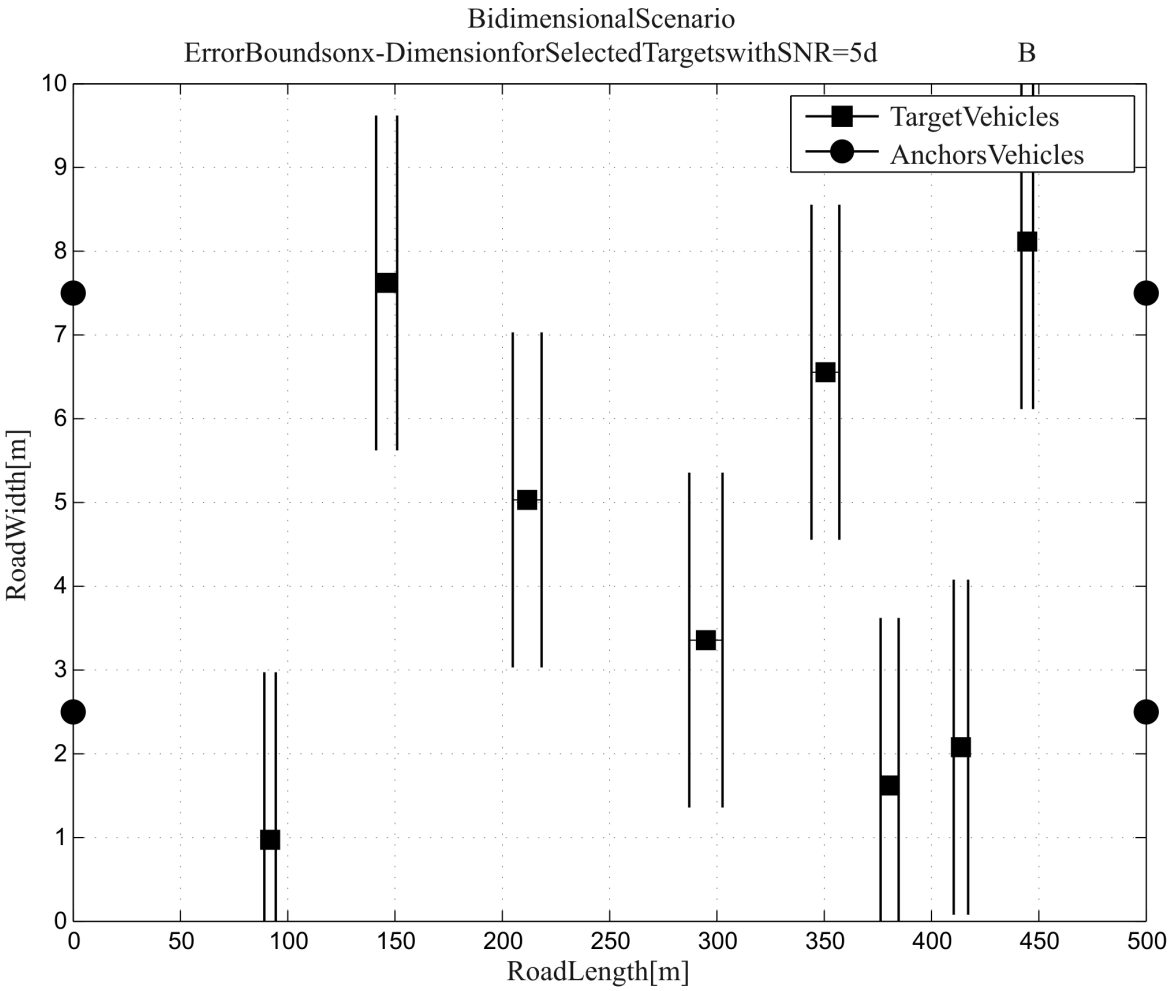


Figure 3: Error Variance for the Bidimensional Scenario. X- and Y-axis are not following the same scale.

available at all. Only a few vehicles, at the beginning and the end of this road, are assumed to correctly self-estimate their own position via GPS. Then we are interested in assessing how much the uncertainty on the GPS position estimates propagates to the rest of the network, and which level of accuracy is possible to reach in anchor vehicle localization.

In order to address the first goal, and therefore to focus only on the impact of GPS accuracy, we have considered a mono dimensional scenario. This can be done without any loss of generality for two reasons: first because the width of the road can be ignored as its ratio to the length is much smaller than one, and second because the road is already a bound for vehicles in the dropped dimension (i.e. cars are supposed not to be out of track). As can be seen from eq. (18), the anchor vehicle position uncertainty effects the (FIM) and hence the (CLRB) of the individual targets.

We initially compare the performance limits of the *accurate* (i.e. no GPS error) versus *inaccurate* ($\Sigma_k > 0$) anchor position estimates. Then we consider the effect of (SNR) on the error bounds of the estimated target positions. We deployed 51 uniformly spaced cars inside the road: the first and the last one, laying at the border of the dense environment, referred

to as anchor vehicles, are able to successfully employ GPS localization system. The error standard deviation (computed according to sec. 3 but for monodimensional case) on the position estimate of the remaining 49 GPS-blinded vehicles, referred to as targets, is then used as metric to evaluate the accuracy of the proposed multihop scheme.

In fig. 1, different values of Σ_k are considered, with a reference (SNR) of 5 dB for 1 meter path and $\alpha = 1.8$ [Pai08]. Clearly the GPS error does not strongly propagate within the network; in particular, for the target at the centre of the road, the standard deviation for $\Sigma_k = 0.9$ is almost the same as the case without anchor uncertainty. In fig. 2 we have evaluated the accuracy for different values of (SNR). With only 10 dB the mean error, in the worst case (for the vehicles at centre of the road) is less than 2 meters; surprisingly, even for poor (SNR) values (such as 0 dB) a quite accurate localization is still possible.

In order to prove the validity of the proposed scheme, we are then back to the bidimentional case scenario, where we consider the same 500 meters road (x-dimension) with a width of 10 meters (y-dimension). We assumed that vehicle positions along the x-dimension road follow a Poisson distribution with $\lambda = 49$, while on the y-dimension they follow a bivariate gaussian distribution, placing them in two different lanes ($\mu_1 = 2.5, \mu_2 = 7.5$) with unitary variance. We have now 4 anchor vehicles, two at each border of the lane; two vehicles are considered connected if their mutual direct distance is less than 70 meters, creating the corresponding neighbourhood set for each target. Each vehicle propagates its relative distance set to its neighbours to the whole network and consequently computes the shortest path, in term of lowest $\bar{\sigma}_k^2$, to each anchor vehicle. For each target vehicle we then determined the corresponding 95%-confidence error ellipse and we used the resulting semiaxis along the x-dimension to asses the error interval around their true positions. For sake of clarity, we then plotted the only a sample of 8 vehicles along the two lanes: fig. 3 shows that even for targets quite far away from anchors, the dispersion along the x-dimension is small enough to make reliable position estimates.

5 Conclusion

We provided theoretical evidence to the fact that V2V multihop communication schemes can be used to localize, with a satisfying precision, mobile vehicles inside dense urban environment, where GPS signal is typically poorly received. We validated such fundamental localization limits assessing the multihop scheme with a newly formulated version of the (FIM) to compute the final (CRLB). The theoretical evidence corroborates and strengthens similar indications found in a growing body of work in which V2V multihop communication scheme is analyzed to be a practical algorithm in order to estimate the position of any vehicle inside an urban dense environment. As seen from the results (and how it would be clear by simulation in the camera ready version of the paper), under multihop framework, anchor uncertainty has negligible effect on the target error position estimates. This advocates for practical algorithms based on multihop framework.

6 Acknowledgements

This work has been performed within the framework FP7 European Union Project BUTLER (grant no. 287901).

References

[Jou08] D. JOURDAN, D. DARDARI, and M. Z. WIN: "Position Error Bound for UWB Localization in Dense Cluttered Environments". In: *IEEE transactions on aerospace and electronic systems* (2008), pp. 613–628.

[Kay93] S. M. KAY: *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993. ISBN: 0-13-345711-7.

[Nic09] M. NICOLI and D. FONTANELLA: "Fundamental Performance Limits of TOA-Based Cooperative Localization". In: *IEEE International Conference on Communications Workshops*. 2009, pp. 1–5.

[Pai08] A. PAIER, J. KAREDAL, N. CZINK, C. DUMARD, T. ZEMEN, F. TUFVESSON, A. F. MOLISCH, and C. F. MECKLENBRÄUKER: "Characterization of Vehicle-to-Vehicle Radio Channels from Measurements at 5.2 GHz". In: *Wireless Personal Communications* (2008).

[Sev12] S. SEVERI, G. ABREU, J. SALORANTA, and MACAGNANO: "Algebraic Confidence for Sensor Localization". In: *Proc. Asilomar Conf. Signals, Systems, and Computers*. Nov. 2012.

[Tor84] D. TORRIERI: "Statistical Theory of Passive Location Systems". In: *Aerospace and Electronic Systems, IEEE Transactions on* AES-20.2 (Mar. 1984), pp. 183–198. ISSN: 0018-9251. DOI: 10.1109/TAES.1984.310439.

Corresponding author: Alireza Ghods, Stefano Severi, Giuseppe Abreu, Jacobs University, Campus Ring 1, 28759 Bremen, Germany, e-mail: [a.ghods,s.severi,g.abreu]@jacobs-university.de

Simulation of Wave Propagation for Radio and Positioning Planning inside Aircraft Cabins

Julia Ringel[1], Samuel Klipphahn[2], Oliver Michler[2]

[1] Fraunhofer Institute for Transportation and Infrastructure Systems IVI

[2] Technische Universität Dresden

Abstract

Various ITS applications need localization or positioning, such as those for people or different types of objects or equipment. In order to support the planning and evaluation of localization based services, a simulation tool for radio wave propagation will be investigated here. Even though radio wave propagation simulation is a suitable and accepted tool for evaluating different transmitter stations (i.e. access points) and versions of antennas, it is not commonly used for the planning and evaluation of positioning applications. By the example of an ITS use case for an aircraft cabin, we will demonstrate in this paper that radio propagation simulation is also a valuable tool to support positioning application planning.

Keywords: radio wave propagation simulation, multipath conditions, ranging, positioning planning

1 Introduction

Wireless Sensor Networks (WSNs) have gained increasing importance in modern traffic telematics applications during the past years. They enable numerous communication applications like the propagation of sensor data or control of actuators. Beyond that, the field of WSN based localization is growing more and more. The main advantage compared to Global Navigation Satellite Systems (GNSS) lies in their applicability in indoor environments and other situations with poor GNSS availability. Applying WSN in aircraft seems promising as well. A common use case is the monitoring of important elements of the aircraft structure, which is called structural health monitoring (cp. [Yed11], [Oli12]). In addition, various applications inside the aircraft cabin, such as backrest or seat belt monitoring and life vest localization, are possible (cp. [Bac11]). WSN planning for aircraft cabin applications is challenging. Firstly, indoor environments have strong multipath characteristics. Secondly, the costs of timeslots for in-aircraft measurement are very high, so measurements have to be reduced to a minimum.

For this reason, physical aircraft mock-ups are built to perform radio measurements in similar environments. Going one step further, radio propagation simulation programs provide the ability to investigate all questions of coverage and localization with a digital model, including electrical properties of an aircraft cabin. [Bac11], [Hoc08] and [Rie02] focus on developing channel models for radio propagation for different types of aircraft cabins, but no conclusions about ranging or localization are drawn so far. Current projects dealing with this topic are e.g. “Neue Elektronische Luftfahrtsystem Ansätze” (NELA) and “Cool Public Transport Information” (CPTI). Within NELA localization is used to identify positions of different objects, e.g. live vests, inside aircraft cabins. Here radio propagation simulation helps to assess position accuracy. The more general issue of how radio ranging is effected by indoor environments is one objective of CPTI. As the considered environments of public transportation vehicles can serve as examples, the results are also valuable for similar environments like aircraft cabins.

This paper is structured as follows: In the beginning, it will give a brief introduction to modeling and simulation concepts for radio wave propagation in general. In chapter 3, the specific use case of radio wave propagation simulation within an aircraft cabin model will be considered. In it, we will suggest suitable evaluation metrics and discuss the results for this use case from the positioning point of view. This way, we will show that radio wave propagation simulation is a valuable tool to support positioning planning e.g. in the context of an ITS application.

2 Principles of radio wave simulation modeling

2.1 Radio propagation algorithms

Radio propagation algorithms can be classified as empirical, semi-empirical, numerical and ray based approaches [Gen98]. As empirical models are derived from measurements, they are valid only for situations similar to those in which their data base was obtained. Even though some are enhanced by additional sub-models considering physical processes, the resulting semi-empirical models cannot be utilized for arbitrary environments like aircraft cabins. Examples of typical areas of application and descriptions of several approaches are given in [Sau07] and [Rap02]. The most accurate solutions would be obtained by solving the Maxwell’s equations. This class of numerical models requires intense processing power. An overview of the most common approximation methods for this class is given in [Gus06].

Regarding all relevant transmissions as separate paths is an assumption that leads to ray based propagation models. By way of example, [Gen98] depicts that for the numerical Finite Difference Time Domain Method (FDTD) the computational time was 10 times higher than that for the ray based model, while the results showed wide conformance. Compared to empirical models, ray based models can be more flexible if the used software implementation supports individually modeled environments.

Hence, a ray based approach is best suited for the examination of radio wave propagation within an aircraft cabin. As software implementation, the program Radiowave Propagation

Simulator (RPS), version 5.5, by the company Actix GmbH was used [Act11]. Principles of the applied software's components are illustrated by Figure 1.

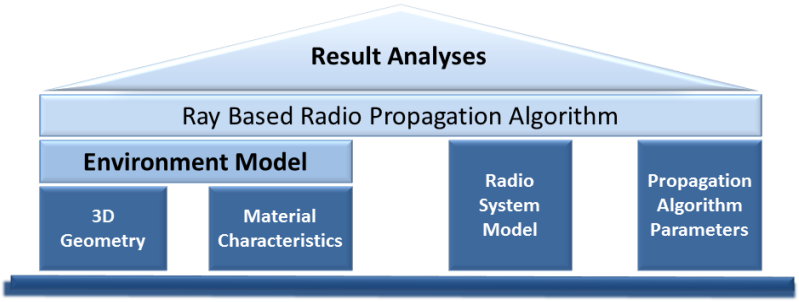


Figure 1: Components of the applied ray based radio propagation simulation model.

The applied simulation tool requires an environment model, a radio system model and propagation algorithm parameters as input data. These components will be described in the two subsequent sections. On this basis, the implemented ray based radio propagation algorithm creates output data for result analyses described in section 2.4.

2.2 Environment model

An environment model for radio wave propagation comprises the radio propagation relevant elements of the environment, such as geometric and electrical characteristics. The detection and computation of multipath effects reflection, refraction, penetration, scattering and diffraction are based on them. The following investigations will be carried out using a CAD model of an Airbus A320-300 aircraft cabin (see Figure 2). It was provided by Airbus Operations GmbH in the aforementioned project NELA.

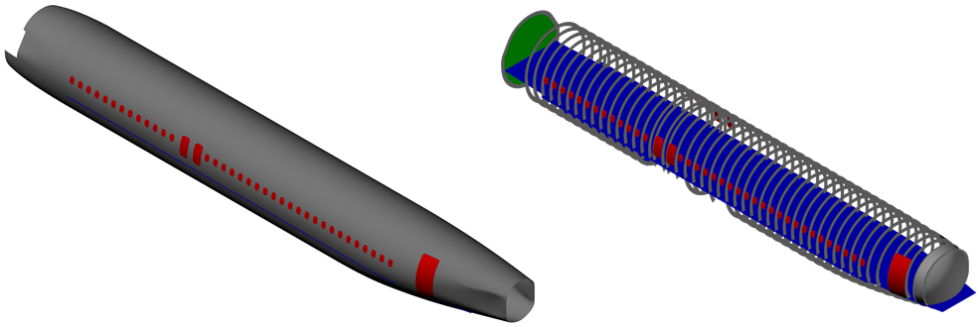


Figure 2: Two different representations of the 3D CAD model of the aircraft cabin

The visualization in Figure 2 contains the complete CAD model, with (left) and without (right) the outer skin. The shown aircraft parts are about 31.50 m in length, 3.80 m in width and have a height of about 4.15 m. The model consists of nearly 456,000 surfaces. These are grouped by material into specific layers of different colors. Besides geometric environment characteristics, the model also comprises material specific electric parameters. The relative complex permittivity ε_r is formed as shown in (1):

$$\varepsilon_r = \varepsilon_r' + j\varepsilon_r'' \tag{1}$$

Hereby ε_r' is the real part, or dielectric constant, and ε_r'' an imaginary part, also termed as loss factor. Table 1 contains permittivity values for specific materials used in material layers of the environment model. These values are valid for the frequency range from 900 MHz to 5.2 GHz.

Table 1: Relative permittivity values of different types of materials and corresponding model layers (material characteristics from [Hoc08] and [Mer08]).

Material	ε_r'	ε_r''	Layers
Aluminum	1	-1000000	cargo floor, doors, pressure bulkhead, stabilizer rings, outer skin
PVC	2.7	-0.1	cabin floor, cockpit wall
Glass	7.76	-0.01167	windows

2.3 Radio system model and propagation algorithm parameters

The radio system model includes all radio communication devices and their characteristics for a given use case. It consists of at least one transmitter and one receiver. The transmitters may represent fixed installed components, the receivers the objects to be localized. Receivers are often modeled as an array or line of individual receiver spots in order to record and plot spatial distributions. In that case, each single spot represents the potential position of an object to be localized. Every transmitter and every receiver has exactly one antenna. Each antenna is characterized by a respective radiation intensity pattern, gain and type of polarization. Different specifics of transmitters and receivers are reached by allocating different antennas. A simplified radio system model is shown in Figure 3.

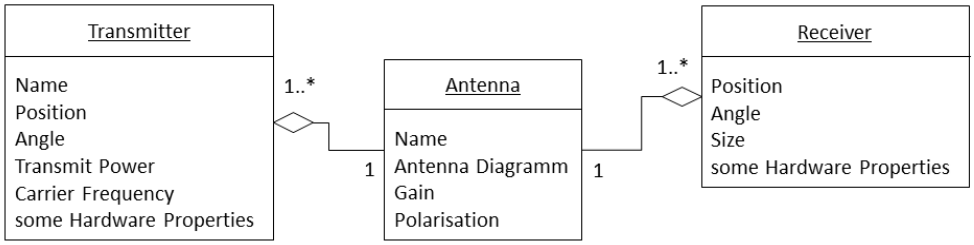


Figure 3: Components of the Radio System Model.

The propagation algorithm parameters define the range of radio wave spread effects within the calculation as well as the termination criteria. Concurrently, these key parameters define the trade-off between processing time and accuracy of calculation results. These settings can be manipulated in every individual simulation experiment.

2.4 Result analyses

The applied software RPS [Act11] offers to analyze results referred to transmitter, receiver with point, path or spatial context. Output options are different diagram types like surface-plots or histograms as well as tables and raw data. They contain diverse metrics like signal

strength, signal runtimes, direction or angle of rays as well as derived quantities like Signal-to-Interference Ratio (SIR) or best serving transmitter.

The decision about the most suitable metrics for evaluation depends on the underlying physical effects of ranging and positioning. There are three basic properties in relation to electromagnetic propagation [Ben08]:

- Received signal strength (RSS): The distance is estimated based on the measured signal strength, typically based on the free space loss equation.
- Angle of arrival (AOA) or direction of arrival (DOA) utilizes the measured direction of the radio wave.
- Time of flight (TOF): Here, a distance is derived from time of flight of the electromagnetic signal. Versions of this ranging principle are time of arrival (TOA), time difference of arrival (TDOA) and phase of arrival (POA).

Within the used software RPS the signal strength delivered by the so called coverage plot. Multipath propagation modifies the angle and the runtime of the received signal. These changes are expressed by plots of angular spread and delay spread. More precisely, the delay spread in RPS corresponds to the total-excess-delay $\Delta\tau$ in equation (Eq. 2), where $\tau_{\min}$ is the signal runtime of the shortest and $\tau_{\max}$ is the signal runtime of the longest ray.

$$\Delta\tau = \tau_{\max} - \tau_{\min} \tag{2}$$

As RPS also supplies all considered rays at each receiver, any delays, as shown in [Bac11], can be received from the Power-Delay-Profile as well. So besides the total-excess-delay, the mean-excess-delay $\overline{\Delta\tau}$ considering the magnitude $P(\tau_i)$ of all incoming rays can be obtained by the difference of the mean and the minimum signal runtime as given in Eq. 3:

$$\overline{\Delta\tau} = \frac{\sum_{i=1}^{N} \tau_i P(\tau_i)}{\sum_{i=1}^{N} P(\tau_i)} - \tau_{\min} \tag{3}$$

3 Use case: ranging in an aircraft cabin

3.1 Specification of the radio system model

3D positioning based on ranging requires a minimum three values between different known positions and the object to be localized. As additional known positions enhance the ranging accuracy, a WSN configuration with six fixed anchor nodes, respective transmitters, is considered. Further assumptions for the radio system are based on the properties of WSN transceivers [Atm13], selected by example and illustrated in Figure 4.

Figure 4: Two WSN transceivers [Atm13a] for ranging.

Even though the modules would provide antenna diversity the model was reduced to one antenna per node due to simulation complexity. All antennas are assumed as dipoles with vertical polarization, a transmit power of 0 dBm and a gain of 2 dBi. Carrier frequencies were set to the center frequency of the medium communication channel of the 2.4 GHz-ISM-Band as 2.44 GHz. Receivers are arranged as an array at a height of 1 m above the cabin floor, following from the height of life vests below passenger seats. The grid size of the receiver field is set to 5 cm resulting from a trade-off between simulation runtime and results resolution. All simulations were calculated with a 2.5D ray tracing algorithm, considering reflections also at non-vertical planes. The number of reflections and penetrations was restricted to five. Diffraction and scattering was disabled because of the expected low impacts compared to the additional computation time.

3.2 Results related to radio planning

Coverage is the typical result parameter in radio planning. It indicates whether the received signal is sufficient. The sensitivity of the radio modules is given in the specification sheet [Atm13] with -88 dBm in the case of maximum data rates. Consequently, each receiver spot must have at least -88 dBm – providing that no interference exists. The focus of this paper is on how to locate the transmitters so as to provide sufficient signal strength at each receiver spot. So for each of the six planned transmitters a single simulation run is required. The objective is to achieve sufficient signal strength for each of the receiver spots in order to be able to identify objects at each possible location.

Figure 5 shows two coverage plots, each gained by a separate simulation run with exactly one active transmitting antenna. Active transmitters are depicted with red circles, all others as blue circles.

The coverage plot indicates suitable signal strength even for the most distant receiver spot. Since there are no obstacles between transmitter and receiver, the distance between them has the highest impact on the result. A similar result was found for the antennas at the opposite end of the aircraft cabin. Even better results were found for antennas in the center section. Furthermore all opposite antennas induced symmetrical coverage likewise the two plots in Figure 5.

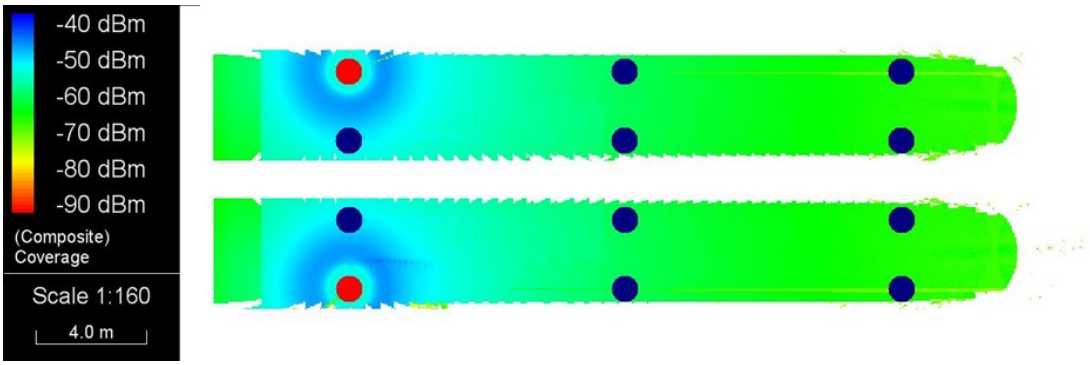


Figure 5: Two Coverage plots of an aircraft cabin (A320-200), - each simulated with one active transmitter in the front section.

3.3 Results related to ranging accuracy

Effects caused by multipath propagation can be identified from delay spread, delay mean spread and from angular spread.

Delay based results

In order to achieve exactly identical color scales, the delay plots in Figure 6 and Figure 7 were created from exportable raw data because the simulation software does not offer a mean spread plot. Figure 6 shows the delay spread plots as described in Eq. 2 for two different transmitters.

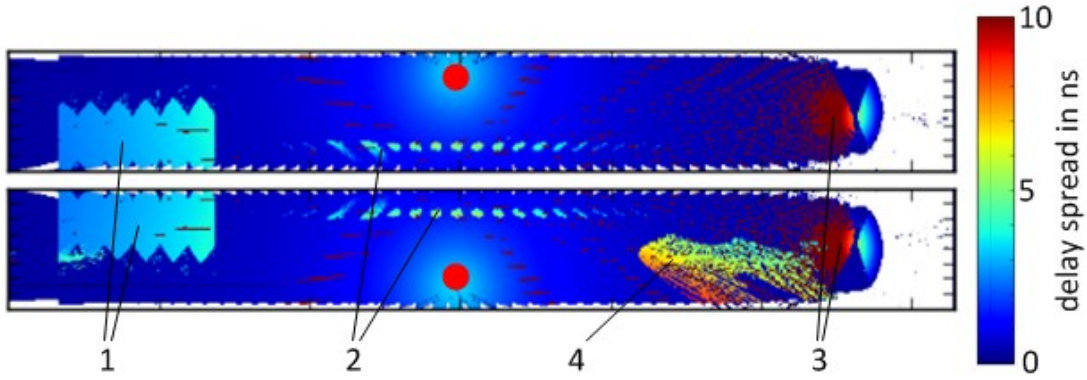


Figure 6: Delay Spread for two different transmitters in the aircraft cabin center.

The increased delay spread values marked with (1) in the front section of Figure 6 originate from reflections at the front cargo room. The shapes are caused by shadowing of the stabilizer rings. Reflections from windows and cabin doors lead to the spots in the aircraft cabin center (area (2) in Figure 6). In the aft section of the right hand side of Figure 6 two different phenomena can be identified. First, the pressure bulkhead acts as a concave mirror causing high delay spread values in area (3). Second, increased delays in the lower figure in area (4) arise from reflections at the rear cargo door, located on the right side. As

this effect is not visible for the antenna on the right side (upper plot in Figure 6) one can conclude that more accurate ranging results are produced here. Hence, in area (4), ranging values from the antenna at the right side should be preferred to those from the left to increase positioning accuracy. A conclusion for the project NELA is that real positioning values for live vests within areas (1), (2) and (4) should be tested particular carefully. Area (3) needs no special consideration as no seats and therefore no life vests are placed there.

The delay mean spread plots, calculated with Eq. 3, are depicted in Figure 7 for the same transmitter locations as in the previous figure.

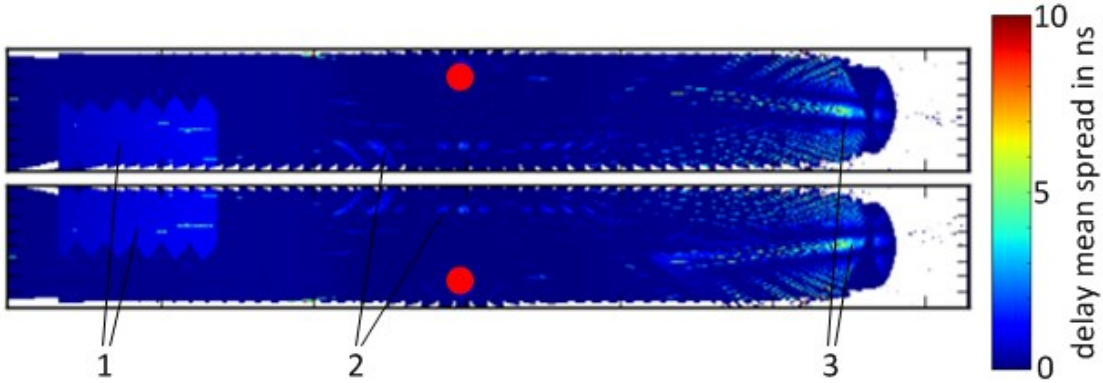


Figure 7: Delay Mean Spread for two different transmitters in the aircraft cabin center.

Similar to Figure 6, Figure 7 also shows increased delay mean spread values caused by reflections of the front cargo room (area (1)), windows, doors opposite the transmitters (2) and the pressure bulkhead (3). In contrast, effects of the rear cargo door (area (4) in Figure 6) are not visible in Figure 7. This is due to the low impact of signals penetrating the cabin floor and being reflected at the cargo door on the composite signal.

Generally, compared with the mean spread in Figure 7, the delay spread in Figure 6 shows a broader range of colors, and respectively higher values. Therefore, the delay spread might be regarded as the more sensitive of the metrics. In exchange, the delay mean spread considers the magnitude of all rays at one receiver and gives better insight into the impact of multipath propagation. Finally, with the help of the propagation speed, the speed of light, one can also infer the achievable ranging accuracy. Disregarding any source of errors from electronic modules or measuring procedures, a resulting signal-run-time increased by 1 ns through multipath effects would lead to a distance measure 30 cm too high.

Results gained from different metrics

As can be seen above, delay based results can help to assess the accuracy of TOA based ranging techniques in different areas of a specific use case. While applying other ranging or positioning techniques, different metrics for accuracy assessment have to be considered as well. Figure 8 shows angular spread plots for the same transmitters.

In comparison with Figure 6 and Figure 7, the angular spread analysis in Figure 8 reveals reflections from the pressure bulkhead in the rear section (area (3)), as well as from windows and cabin doors opposite the transmitter (area (2)), whereas reflections from the

front and the rear cargo room (area (1) and (4) in Figure 6) are not visible. This means that by angular spread analysis, fewer areas influenced by multipath propagation are identified.

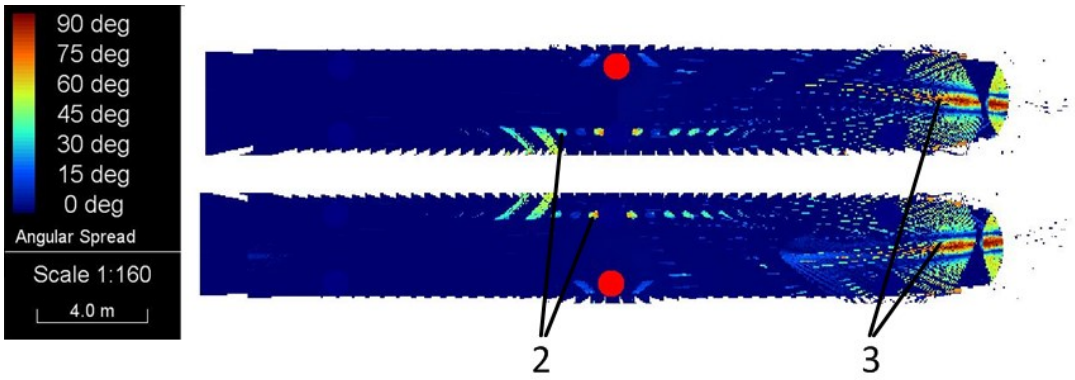


Figure 8: Angular Spread for two different transmitters in the aircraft cabin center.

Figure 9 visualizes how ray paths lead to different results for delay compared to angular spread. The two ray paths between one transmitter and one receiver in front of the center cabin door cause a high angular spread value of about 50 degrees while delay spread values are only about 3 ns. Hence, positioning based on TOF would be more eligible for this area than positioning based on AOA.

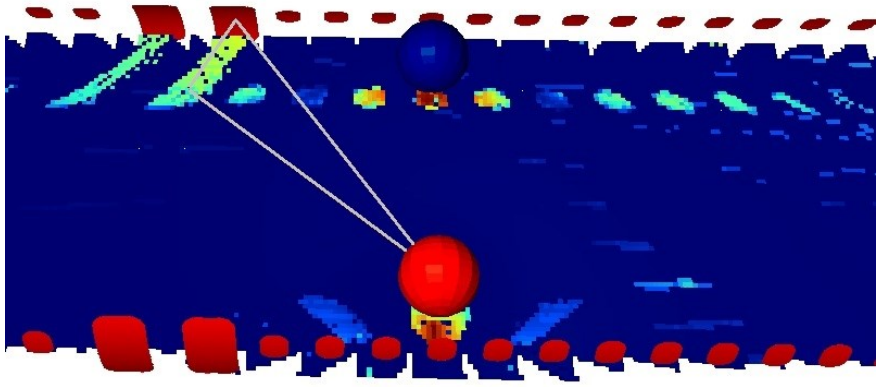


Figure 9: Ray paths from central left transmitter to an exemplary receiver position with reflection at a central cabin door.

Depending on the applied measurement quantity, the occurring areas influenced by multipath propagation are partly different. So if for a given environment a decision about positioning techniques with different underlying physical effects has to be made, radio wave simulation can also support. This procedure can be executed also for other use cases such as for public transport vehicles which were considered in the project CPTI.

4 Conclusions and outlook

4.1 Conclusions

Our exemplary study of an aircraft cabin demonstrates how radio wave propagation simulation is applicable not only for coverage evaluation, but also to identify the influence of

different surroundings on ranging and position accuracy. Similar to radio coverage planning, a simulation approach can be used to find areas that show different levels of ranging accuracy caused by environmental conditions. Comparing the different result parameters can contribute to finding the most useful ranging technology. So, the results gained with radio wave propagation simulation can help to develop locating approaches for ITS applications as in the projects NELA and CPTI.

4.2 Outlook

Future work will have to analyze the received phase compared to the phase to be received without multipath spread. This additional metric is especially valuable for ranging purposes based on POA, e.g. [Atm13]. In addition, simulation results will be compared to measured values for relevant ITS use cases. This can be done for signal level, signal run times and other quantities. On the other hand, a comparison of real ranging accuracies in correlation to environmental effects with respect to receiver and measurement accuracy could be valuable as well.

Acknowledgment

This work was partly supported by the projects NELA and CPTI. NELA is funded by the German Federal Ministry of Economics and Technology, CPTI by the European Union and the Free State of Saxony within the cluster COOL SILICON.

References

[Act11] J. DEIßNER, J. HÜBNER, D. HUNOLD, J.VOIGT, and P. SCHAFFER: *Radiowave Propagation Simulator. User Manual- Version 5.5*. Ed. by ACTIX GMBH

[Atm13] ATMEL CORPORATION: *AT86RF233 Preliminary*. Feb. 2013. URL: http://www.atmel.com/Images/Atmel-8351-MCU_Wireless-AT86RF233_Datasheet.pdf.

[Atm13a] ATMEL CORPORATION: *Atmel AT86RF233 Evaluation Kit*. Sep. 9, 2013 URL: http://store.atmel.com/PartDetail.aspx?q=p:10500317#tc:description.

[Bac11] M. BACHHUBER: "Analyse und Modellierung der Funkausbreitung in Passagierkabinen von Großraumflugzeugen". PhD thesis. Erlangen, Germany: Universität Erlangen, 2011. URL: http://www.opus.ub.uni-erlangen.de/opus/volltexte/2011/2510/pdf/MartinBachhuberDissertation.pdf.

[Ben08] A. BENSKY: *Wireless Positioning Technologies and Applications.* Artech House, 2008.

[Hoc08] N. HOCKE: "Systematische Untersuchungen zur Steigerung der Recheneffizienz beim Funkplanungstool RPS für ausgewählte Vielsenderszenarien und

dynamische Simulationsumgebungen". diploma thesis. Dresden, Germany: HTW Dresden, 2008.

[Gen98] N. GENG and W. WIESBECK: *Planungsmethoden für die Mobilfunkkommunikation*. 1. Edition. Berlin, Heidelberg, Germany: Springer, 1998.

[Gus06] F. GUSTRAU and D. MANTEUFFEL: *EM Modeling of Antennas and RF Components for Wireless Communication Systems*. 1. Edition. Berlin, Heidelberg, Germany: Springer, 2006.

[Mer08] M. MERKEL: *Taschenbuch der Werkstoffe*. Leipzig, Germany: Carl Hanser, 2008.

[Oli12] L. M. L. OLIVEIRA, J. J. P. C. RODRIGUES, B. M. MAÇÃO, P. A. NICOLAU, and L. ZHOU: "A WSN Solution for Light Aircraft Pilot Health Monitoring". In: *Wireless Communications and Networking Conference: PHY and Fundamentals (WCNC), IEEE*. Vol. 1. Paris, France, 2012, pp. 119–124. ISBN: 978-1-4673-0436-8. URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6213959.

[Rap02] T. S. RAPPAPORT: *Wireless Communications, principles and practice*. 2. Edition. Prentice-Hall, 2002.

[Rie02] N. RIERA and M. HOLZBOCK: "Aircraft Cabin Propagation for Multimedia Communications". In: *Proceedings of the 5th European Mobile & Personal Satellite Communications Workshop (EMPS 2002).* Baveno, Italy, 2002, pp. 281–288. URL: http://www.triagnosys.com/wcab/download/papers/EMPS2002.pdf.

[Sau07] S. R. SAUNDERS and A. ALEJANDRO-ZAVALA. *Antennas and propagation for wireless communication systems*. 2. Edition. Chichester, England: Wiley, 2007.

[Yed11] R. K. YEDAVALLI and R. K. BELAPURKAR: "Application of wireless sensor networks to aircraft control and health management systems". In: *Journal of Control Theory and Applications* 9 (Feb. 2011), pp. 28–33. ISSN: 1993-0623. URL: http://link.springer.com/article/10.1007-s11768-011-0242-9.

Corresponding author: Julia Ringel, Fraunhofer Institute for Transportation and Infrastructure Systems IVI, 01069 Dresden, Germany, phone: +49 351 4640626, e-mail: julia.ringel@ivi.fraunhofer.de

Network-wide application of Floating Car Data (FCD) Particularly in Cities Using Data Fusion with Measurement Data of Existing Stationary Traffic Detection

Ralf Kohlen
VMZ Berlin Betreibergesellschaft mbH

Abstract

An important task of traffic management systems is the generation of a current traffic situation map. Therefore stationary traffic detectors are often installed and operated at strategically relevant locations of the road network. In most cases, for all other streets the traffic situation is being calculated by a traffic model. The resulting quality of those online models and the expenses to operate and to maintain the models and the input data often not fit the operator's requirements.

This paper shows an alternative solution for the city use case as well as for the application outside the cities. It is based on a data fusion method. It combines travel times measured by floating car data (FCD), which are available by now extensively, with measurement data of the existing stationary traffic detection. The method starts with the definition of quality requirements on the resulting traffic situation map. Using the example of Berlin Traffic Information Centre (TIC) the process to implement this method in a real environment and the reached output quality will be shown.

This method to calculate the network-wide current traffic situation produces more realistic results than it was possible using a traditional traffic model in the past. Especially unforeseeable events like accidents or the results of parking cars in second lane will be respected by the FCD data source. Furthermore the data fusion process enables a consistent picture of the current traffic situation, at street sections with existing stationary traffic detection as well as at those without.

Keywords: traffic management, traffic situation, cities, motorways, FCD, floating car data, data fusion, traffic model, traffic detection, quality, coverage, resolution, actuality

1 Introduction

The VMZ Berlin Betreibergesellschaft mbH (hereafter referred to as “VMZ Berlin”) operates the Berlin Traffic Information Centre (TIC) on behalf of the State of Berlin. VMZ Berlin is a SIEMENS subsidiary and operates this former traffic management centre since the year of 2001.

In the year of 2010 the State of Berlin published a tender for the operation of the new Berlin TIC until the end of 2020. The traffic situation map is a kernel module of this centre. A lot of applications use this data. Therefore the State of Berlin has defined a set of special quality requirements on the traffic situation map. These requirements are based on the experiences with the legacy system.

- First requirement: The coverage of the new traffic situation map has to be extended from approx. 900 km to approx. 1.600 km without installing more stationary traffic detectors.

- Second requirement: The actuality of traffic information has to be improved. To reduce latency, the update interval should be reduced from 15 minutes to approx. 5 minutes.

- Third requirement: The quality of the traffic information should fit the reality much better. There are defined threshold values.

The State of Berlin has not forced requirements on the method to calculate the current traffic situation, but on the results. Therefore it was possible to continue operating the existing online traffic model as well as to look for advanced technologies. VMZ Berlin has decided to make a pre-commitment on a special system. In fact there has been an extensive analysis of solutions for this problem being now available on the market, e. g. macroscopic, mesoscopic and microscopic models as well as FCD sources. In the result the decision has been made in favour of a data fusion method described in the following sections.

2 Objectives

Based on the experiences of the existing traffic model used to calculate the current traffic situation, the State of Berlin has defined requirements on the results of the new Berlin TIC system as described above. In the past VMZ Berlin used a macroscopic traffic model (called MONET resp. VISUM online). At the time of its implementation in 2002 the objectives have been focused on the modelling of known traffic issues like road works. But with increasing experiences in live operation the requirements have been increasing as well:

- The network coverage should be extended from a road network of about 900 km to approx. 1.600 km. This requirement follows the Urban Traffic Development Plan of the State of Berlin [Sen11] with the definition of road classes in this document. The challenge of this requirement is to double the covered network without installing new stationary traffic detectors.

- The network resolution has to be increased. That means, the street segments of the digital network model have to be defined in a way, so that there is no traffic lights controlled junction inside of the segments. At least every junction with traffic lights defines an end of a street segment. This is an advantage for the presentation of the traffic map, but it is an additional challenge for the technology, because it increases the number of nodes and links in this model in a significant way.

- The legacy system has calculated the current traffic situation every 15 minutes. This period is too long for most applications, especially for traffic control. Therefore the refresh period has to be reduced, in best case to an interval of about 5 minutes. This has to be respected as well as the network coverage extension as described above.

- The results shown in the traffic situation map should fit the reality much better than before. In 70 per cent of all cases the calculated level of service (LOS, three levels: free flow, slow-moving traffic, congestion) should fit the reality exactly. In 20 per cent of all cases a deviation of one level will be tolerated. These are high-level requirements on the technology, because they are focused on the result, regardless whether the reasons for traffic disturbances are known or not. For instance, the traffic effects of blocked roads due to demonstrations, accidents or tunnel closures should be reflected.

The experiences of the operator have shown high costs for the set-up and for the maintenance of the traffic model, including the network models as well as the demand matrices for 7 days a week and 24 hours a day. And last but not least, the calculation of LOS values for traffic jams due to road works etc. did not fit the requirements of the operator nor those of the user.

3 Evaluation of current methods and models

VMZ Berlin has evaluated different technologies available on the market by now. This includes macroscopic, microscopic as well as mesoscopic traffic models. Other data sources like floating car data and data fusion approaches have been reviewed. The criteria have been as follows:

- Quality of the results: Can the model guarantee, that the results will fit the quality requirements as defined by the State of Berlin?

- Requirements on the input data: What kind of data are required and what are the requirements on the actuality and on the quality of that data?

- Calculation time: How long will it take to calculate the current traffic situation if there is an update of the input data available?

- Costs: How much is it to set up the system and to operate and to maintain it for a period of about 10 years?

This evaluation resulted in the fining that every model has the disadvantage of being not

able to respect missing input data – regardless of the kind and of the quality of the model. But a traffic information centre often has no information about the reason of traffic jams. In live operation this happens very often. Therefore a direct measurement of the traffic effect has real advantages for the application in a real traffic information centre.

The only open questions are about the price and how to fuse the FCD with the local data in an way, that there is no conflict between both data sources. All these questions have been answered by VMZ Berlin for the use case of the new Berlin TIC.

4 Brief description of the new method

The concept of the new data fusion method is based on using pre-processed floating car data in terms of current road segment speeds. Therefore VMZ Berlin uses HD Flow data by TomTom for a network of about 1.600 km in Berlin and, by now, of about 400 km in the surrounding State of Brandenburg (see Figure 1).

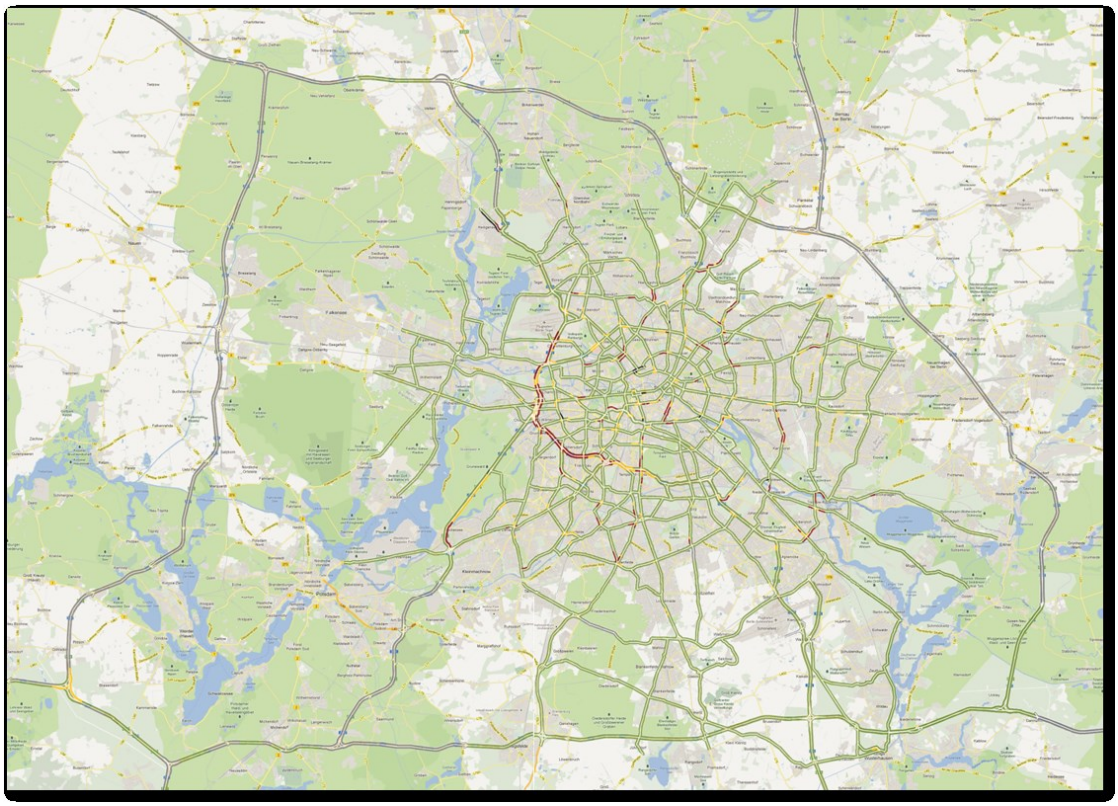


Figure 1: Network coverage in Berlin TIC.

The new fusion technology interprets the FCD in a way that there is no conflict with the LOS information of the local detectors (approx. 1.000). And it respects traffic news like road works, tunnel closures or accidents.

The internal update rate is about 15 seconds, the external rate on the VIZ website (see www.viz-info.de) is about 5 minutes.

5 Results of application in Berlin Traffic Information Centre

The new traffic situation map has been tested in cooperation with both contraction authorities: The State of Berlin and the State of Brandenburg. At august 14th, 2012 the system has been set live on the Berlin TIC's website.

Beginning with the going live the quality of the results have been monitored continuously. The method is to compare the LOS values at all local detectors with the results before the data fusion process. The results fit the requirements not only in all cases of the reporting quarter. They fit the requirements as well in the morning rush hour and in the evening rush hour (see Figure 2)

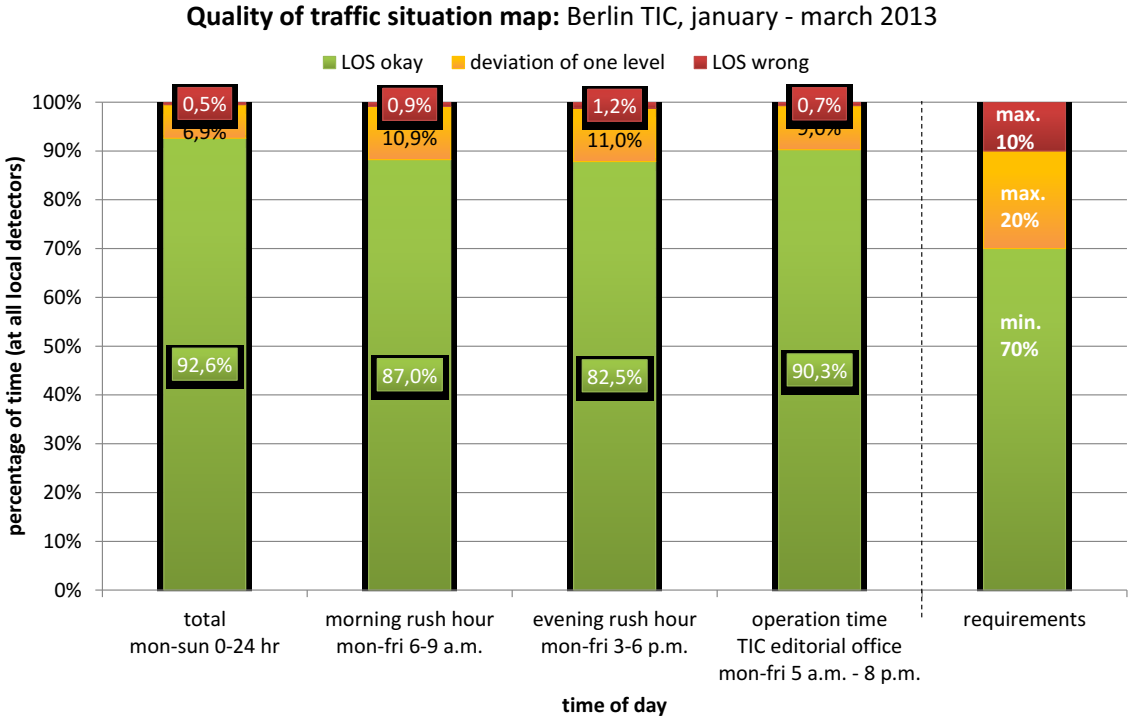


Figure 2: Quality report (example).

6 Conclusions

VMZ Berlin has decided to use da new data fusion method for FCD and local data for calculation of the current traffic situation in Berlin TIC instead of using a traffic model. Because of missing input data concerning reasons for traffic jams the results of no model can fit the requirements. This is the main advantage of using FCD. The traffic effects of known events like road works and unknown events like accidents will be monitored. The new method ensures, that the LOS of the data fusion will fit the LOS measured at local detectors in a high degree within the quality requirements.

References

[Sen11] SENATSVERWALTUNG FÜR STADTENTWICKLUNG UND UMWELT: "Stadtentwicklungsplan Verkehr (StEP)". Berlin, 2011. URL: http://www.stadtentwicklung.berlin.de/verkehr/politik_planung/step_verkehr/

Corresponding author: Dr.-Ing. Ralf Kohlen, VMZ Berlin Betreibergesellschaft mbH, Ullsteinstraße 114, Turm C, 12109 Berlin, Germany, phone: +49 30 81453 125, e-mail: ralf.kohlen@vmzberlin.com

New Challenges in FCD Research

Günter Kuhns, Elmar Brockfeld, Thorsten Neumann, Alexander Sohr, Louis Touko
German Aerospace Center (DLR)

Abstract

While algorithms to process FCD from raw spatio-temporal data became common and more sophisticated during the last years, these data are already widely used by applications for operational traffic management or by traffic data providers which are mostly based only on travel or delay times. To improve the quality of FCD and deduce further traffic parameters new research is conducted, which also creates new use cases for these data. Based on new devices (e.g. smartphones) and lower cost for wireless communication new data sources are available whose different characteristics require adaptions or even different algorithms for processing. This paper will show algorithms to assess and improve quality of FCD by including additional information as well as handling new data sources or extracting additional information for new FCD use cases.

Keywords: FCD, Quality, Applications, Network Models, Fundamental Diagram

1 Introduction

To assess the current traffic situation, predict further developments and react accordingly, operational traffic management requires a good coverage of the road network with accurate real-time traffic information. Sources of these are usually induction loops, roadside cameras, automatic vehicle identification (AVI) systems and lately also Floating Car Data (FCD).

Due to high installation and maintenance costs the deployment of dedicated static sensors for whole road networks is in most cases not feasible. In contrast collection of FCD or similar data is very cost-effective since it uses existing infrastructure (GNSS / mobile communication networks), does not require special devices or sensors (only GPS enabled smartphones) or uses already installed systems (e.g. in taxi fleets) to collect spatio-temporal data and thus constitutes an attractive source of traffic information.

While other systems store information that allows identification of individual drivers (e.g. AVI), FCD only requires the re-identification of a vehicle during one trip and allows collection of traffic data while respecting the privacy of its users. DLR has running FCD systems for several years now and current work is mostly focused on assessing and improving quality of generated FCD results, to develop new sources of traffic data with similar characteristics and on creating new fields of FCD applications.

2 Improving quality of FCD

While first reactions on FCD as a new source for traffic information where sceptical at best, during the last years it became a valuable data source for traffic providers and end users. This increase in trust is based on improvements of algorithms for FCD processing as well as successful measurement campaigns conducted to assess its quality [Bro07]. But based on characteristics of FCD and the raw positional data it is computed from, at least two sources for errors remain: low sampling rates of raw data and spatial errors due to inaccuracy of GNSS.

Due to low sampling rates routes driven by vehicles between subsequent measurements could not be determined unambiguously, especially if not the optimum route but a detour was used. Also the time between two samples has to be distributed on road segments as a combination of travel and waiting times which becomes ambiguous for low sampling rates. Spatial errors can also affect the distribution of travel times but in addition cause errors in the map matching process, if positions are mapped to wrong street segments. This would cause detours which were not part of original routes driven by affected vehicles and result in wrong travel times.

To improve quality of FCD further information can be integrated into the processing of spatio-temporal data or used for alternative methods of quality assessment – two examples to be presented here are historic data and also turn relations of vehicles.

2.1 Self-Evaluation approach

While measurement campaigns are often costly, based on small fleets which can only cover a limited area, the Self-Evaluation approach uses measurements conducted by FCD fleets to assess the quality of algorithms used to process these data. For each vehicle trip the route with its travel time is used as ground truth, since it can be measured with sufficient confidence and wrong distribution of travel times between road segments can be mostly neutralized for a route as a whole. Route travel times are then compared with the sum of travel times on each link which are generated in the same way as the results of the FCD system.

In a first analysis [Kuh11] historic speeds were included in the comparison, since they have high influence on results if there are not enough current measurements. In general travel times were underestimated by the FCD system and since this was even more significant in comparison with historic values, these were identified as the most likely cause for this deviation. The systematic error was highest at times with a high variability of travel times (e.g. morning or afternoon peaks), when also the granularity of historic travel times was not detailed enough to reproduce rapid changes in traffic situation. One possible way to deal with this problem is introduced in the following section.

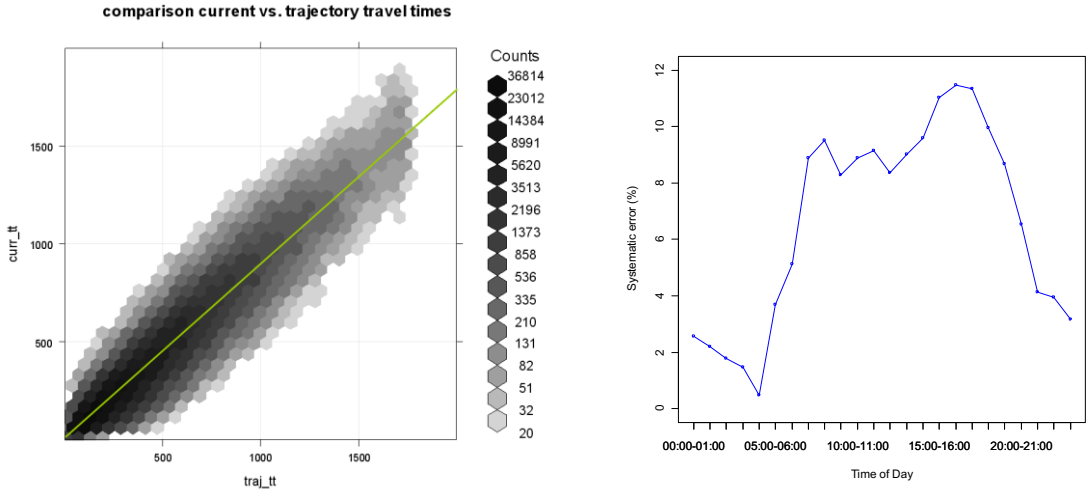


Figure 1: Comparison of travel times (x - trajectory vs. y - system) and average course of systematic error during the day.

Since trajectories which are used as ground truth in this approach are based on raw measurements and are affected only by map matching, most parts of FCD algorithms – like underlying models, distribution of travel times or additional values included in the result - can be tested by this approach to evaluate positive or negative effects on the results. Detailed analyses (e.g. by street type, time of day, area) can identify other specific influences on FCD quality.

2.2 Historic speeds based on Lomb Periodograms

Since travel times are usually similar for same times or weekdays, historic speeds that represent the typical traffic situation for a given time on a road segment are valuable information for predictions. Especially on minor roads FCD are usually not delivered at regular intervals, so historic speeds can be used here to fill gaps in data, if no measurements are available or to smoothen fluctuating or noisy FCD values.

While one of the easiest ways to represent historic speeds are tables of speed values for an edge at given intervals (e.g. over the course of a week), a more elegant and memory efficient way is to use Lomb Periodograms. Here the development of speeds for a given day is represented by a simple trigonometric function with only a few parameters. Thus less information is required (parameters of the function vs. values for all intervals) and values can be stored in a continuous way that is not restricted by interval sizes.

As a first step of the algorithm developed by DLR [Soh09] the raw data for each link is analysed. To ensure that the typical behaviour is represented, data of all public- and school-holidays were rejected and only links in the road network with a significant amount of data are used. To avoid the influence of seasonal changes, the used period of time is limited to the last 3 months.

Based on that pre-filtered data for each used link in the road network and each day of week, the power spectrum of the speed distribution is estimated (curve in Fig. 2b). Due to irregular time intervals, we use a method invented by Lomb [Lom76, Pre07], which is quite

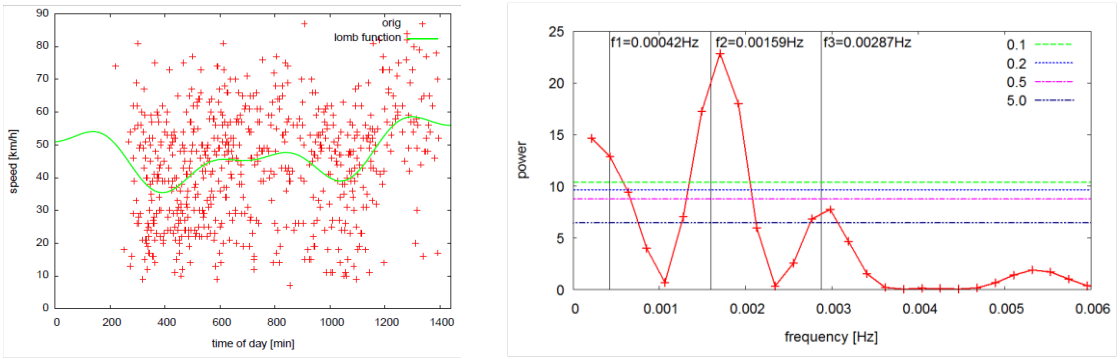


Figure 2: a) Raw FCD with Lomb function and b) power spectrum with significance lines.

different from the normal fast Fourier transformation. Lomb's method delivers, in addition to the power spectrum, also significance levels (horizontal lines in Fig. 2b for 0.1%, 0.2%, 0.5% and 5%) of the frequencies, which is then used to select the most significant ones (vertical lines in Fig. 2b).

These are used in a subsequent step to compute the resulting daily course function:

$$v(t) = v_{0+} \sum_{i=1}^{N} a_i \sin \omega_i t + b_i \cos \omega_i t \qquad (1)$$

Here N is the number of chosen frequencies, $\omega_i = 2\pi f_i$, and $f_1, \ldots, f_N$ are the chosen frequencies. Missing parameters are:

- v_0 – base speed level (50.1 km/h in the example at Fig. 3)
- $a_1, \ldots, a_N$ and b, … , b_N- amplitudes of the trigonometric functions

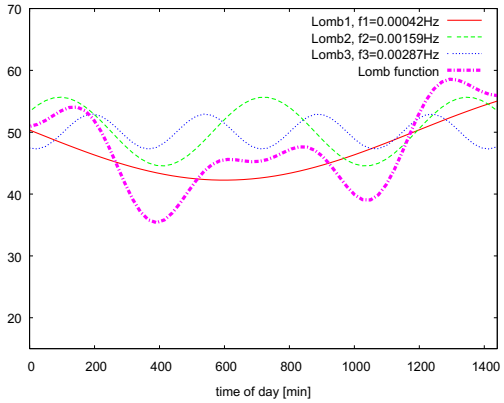


Figure 3: Daily curve (purple) composed from three trigonometric functions.

These parameters can be fitted in one step with singular value decomposition fit (SVD fit), which produces for an over-determined system a solution which is the best in least-squares sense [PVT+07]. The resulting sine and cosine functions are shown in Fig 3. Summed up, these functions give one function (purple) to describe the daily curve of a day with a clear morning and afternoon breakdown, a comeback in the evening and over midday.

A first positive side-effect is the high memory-efficiency of the proposed speed profiles: since the used periodic functions are always sums of sines and cosines, all that has to be saved is v_0, the base speed level, chosen frequencies $f_1, \ldots, f_N$, and parameters $a_1, \ldots, a_N$ and $b, \ldots, b_N$ as the amplitudes of used trigonometric functions. A second advantage is the ability of Lomb-based speed profiles to serve as basis for arbitrary granularities beyond usual day-of-the-week, hour-of-the-day granularity of conventional speed profiles. This is possible since they are a description of speed profiles by means of periodic functions, which of course allows for evaluation in arbitrary intervals.

2.3 Turn dependent travel times

Depending on intersection geometry, concurrent traffic flows, traffic signals and congestion level different turning relations for intersections often have varying delays. While FCD processing usually yields one delay time per intersection inflow, which is the average value of the delay times, it is also possible to extract and analyse this information separated by turning relations. This can be done either by analysing trajectories by inflow-outflow relations of intersections after processing or using a more granular network model that incorporates and separates different turning relations during FCD processing by design.

A first research was conducted for several junctions with high density of measurements in Berlin [Bro10] where DLR has the biggest fleet of FCD vehicles. Each trajectory passing one of these junctions was analysed by its inflow / outflow relation and delay times were determined over the course of normal working days. While in most cases the expected result of the lowest delay for the straight direction could be confirmed, in some cases it was even reverted especially in highly congested traffic situations or if another relation was privileged by traffic signals.

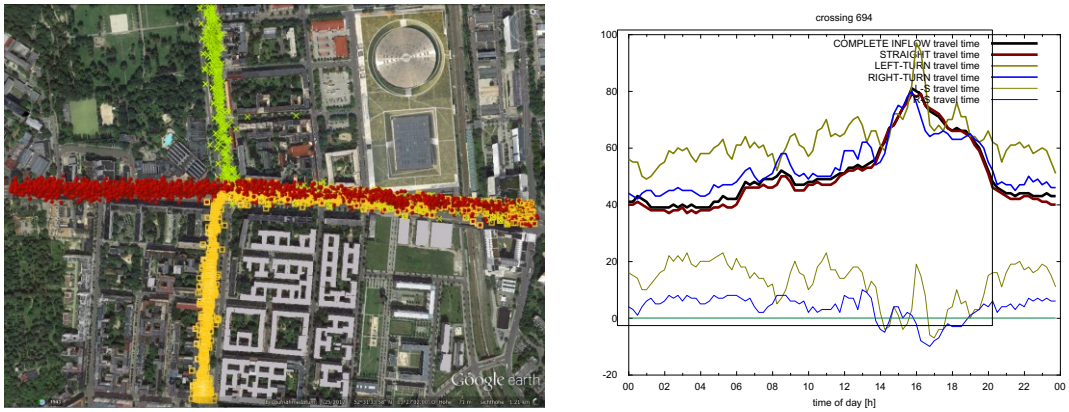


Figure 4: Separation of intersection inflows and delay times per turn relations.

While these results are generated by post-processing and are only available for a limited set of selected junctions a more general approach was incorporated into FCD processing itself which only requires an automated pre-processing of junctions to determine possible inflow/outflow relations. Based on the results of this pre-processing the base network is extended accordingly with new separate inflow links for each turning relation, that are

connected only with one specified outflow. The FCD algorithm will then generate for all vehicles routings on respective inflow links, which are associated with the used turning relations and also map travel or delay times on those.

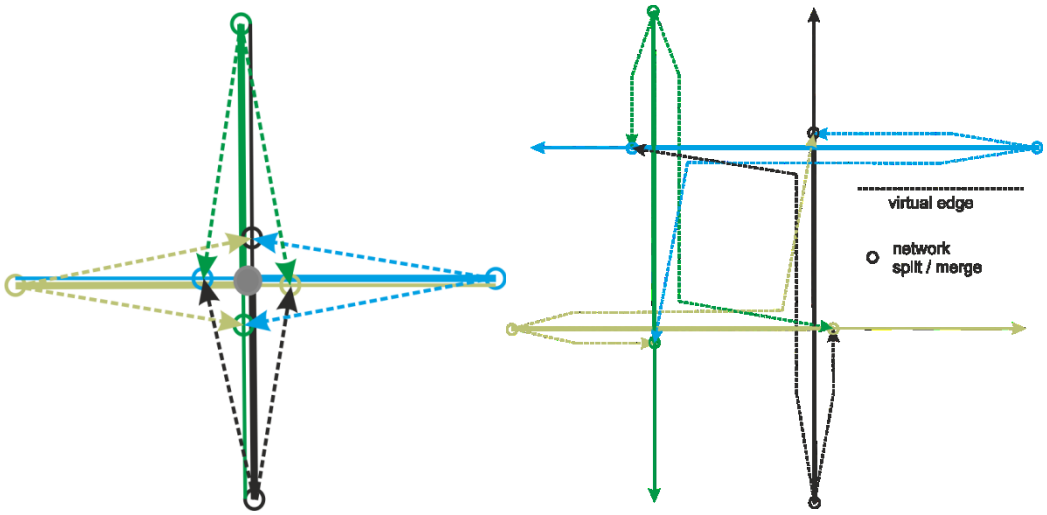


Figure 5: Network model with separate intersection inflows (simple and complex case).

3 New Data sources and applications

As generation of traffic data from Floating Cars became an accepted source of traffic information also similar sources of spatio-temporal data are being developed lately by tracking mobile devices like GPS enabled smartphones. As classic FCD is usually based on data from devices that are connected with vehicles of a fleet, vehicle type and traffic mode of measurements were fixed, which usually does not apply for these new data sources. Smartphones could for example also record traces of pedestrians, so the traffic mode has to be transmitted or determined from collected data, to filter results before they are used for further applications.

Lower costs for communication allows the transfer of data with higher sampling rates or additional information like local speeds or bearings. By using this information for FCD processing more precise results can be generated.

3.1 Mode detection

If measurements from different traffic modes are mixed much valuable information is lost, which can be preserved if measurements are filtered by traffic mode. The most important information to determine the traffic mode are speed or acceleration, which can be detected either directly by sensors of smartphones or derived from subsequent spatio-temporal measurements. According to the course of trajectories some modes could be excluded, but also additional less obvious information like charging can be used. If it is feasible (privacy issues) to generate profiles for certain devices, an assumption about the traffic mode can also be made from past behaviour.

For selected roads (e.g. highways) where measurements can only be caused by vehicles

data collected from GPS enabled smartphones could be used without mode detection for a high detailed traffic monitoring [Her09].

3.2 High frequency FCD

With higher sampling rates the travel or delay times can be determined with a higher accuracy and better assignment to road segments or intersections. Also the progress of trajectories or accumulations of measurements at congested areas or before traffic signals can be determined with sufficient accuracy for new applications (e.g. determine acceleration or length of traffic jams) which were not possible with currently usual sampling rates for FCD fleets. Most data used by DLR so far has sampling rates in the range between 10 to 120 seconds, but due to the reasons named above, we expect to see FCD with higher sampling rates (e.g. 1 to 10 seconds) more often in the future and have already collected and evaluated some of these during measurement campaigns.

Due to higher proximity of subsequent measurements the influence of GPS-errors or deviations becomes more significant compared to errors caused by low sampling rates. If erroneous positional measurements are assigned to wrong road segments, this will cause virtual detours in vehicle trajectories resulting in wrong or too low travel times for the affected street segments.

An approach that could be used before map matching would be “spatial smoothing” – since trajectories are often curves, values that do not fit into the current curve could be corrected or ignored, since they are often caused by GPS-errors. But this could also remove trajectories which are caused by abnormal driving behaviour (e.g. abrupt change in direction).

After positions are mapped on a street network and the route of a vehicle is computed, speeds on this route can be checked for plausibility. This approach can fix detours caused by wrong map matching (e.g. position mapped on a nearby lane with opposite direction) but could also filter out real detours. Since this is also feasible for FCD with lower sampling rate that plausibility check is also used by the FCD algorithm of DLR.

A comparison of speeds between subsequent measurements on a route and smoothing of speeds could fix GPS-errors in or against current driving direction, but would also filter out rapid acceleration or braking.

Except for the second one these approaches work only with FCD of higher sampling rates and are part of further research resulting in FCD of higher quality based on the benefits of a higher sampling rate.

3.3 Deriving traffic flows from FCD using the Van Aerde model

Traffic volumes are one of the most important figures for operational traffic management, infrastructural planning and traffic models or simulations. Due to high infrastructure costs it is currently not feasible to measure those on the network level. Instead models are used to

interpolate these values from punctual measurements of traffic flows at selected points of the road network [Vor06, Hen 07].

While FCD does only measure travel speeds, it is available area wide. Based on travel speeds from FCD combined with a Van Aerde model [VAe95], that uses the speed-flow relationship of the fundamental diagram, it is also possible to determine traffic flows, not only at certain locations but on the network level.

This model is based on the assumption that the headway $h(v)$ of vehicles is based on the free flow speed v_f and a current speed $v < v_f$ together with suitable variables c_1, c_2 and c_3 which are specific for a road class. By $d(v) = \frac{1}{h(v)}$ and $q(v) = v\, d(v)$ the original model can be used to show the relation between current speed v and flow $q(v)$:

$$q(v) = \frac{v}{c_1 + \frac{c_2}{v_f - v} + c_3 v} \qquad (2)$$

The first step is to calibrate Van Aerde models for different road classes, that are specified by speed limit, number of lanes and road class (e.g. highway, minor road) and for which measurements of the speed and traffic flow are available. For this the deviation of measurements and the Van Aerde curve has to be minimized.

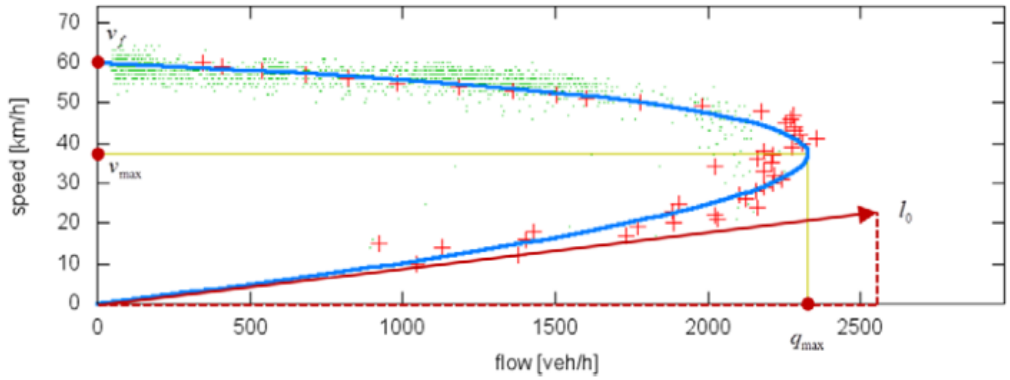


Figure 6: Model calibration.

The resulting curve (Fig. 6) is then used to determine v_f, maximum capacity q_{max}, speed at maximum capacity v_{max} and l_0 as gross vehicle headway for jammed traffic, which define the model parameters for a road segment. Based on the Van Aerde model, traffic volumes can be computed from FCD speeds also for road segments with the same or similar characteristics for which no direct measurements of traffic flows are available.

This method was applied for the city of Berlin where DLR had access to data from induction loops and also FCD from about 4,300 taxis to calibrate Van Aerde models for 25 road classes [Neu13b]. Detectors which were not used for calibration were used as reference sensors to evaluate traffic flow results based on these models and FCD speeds.

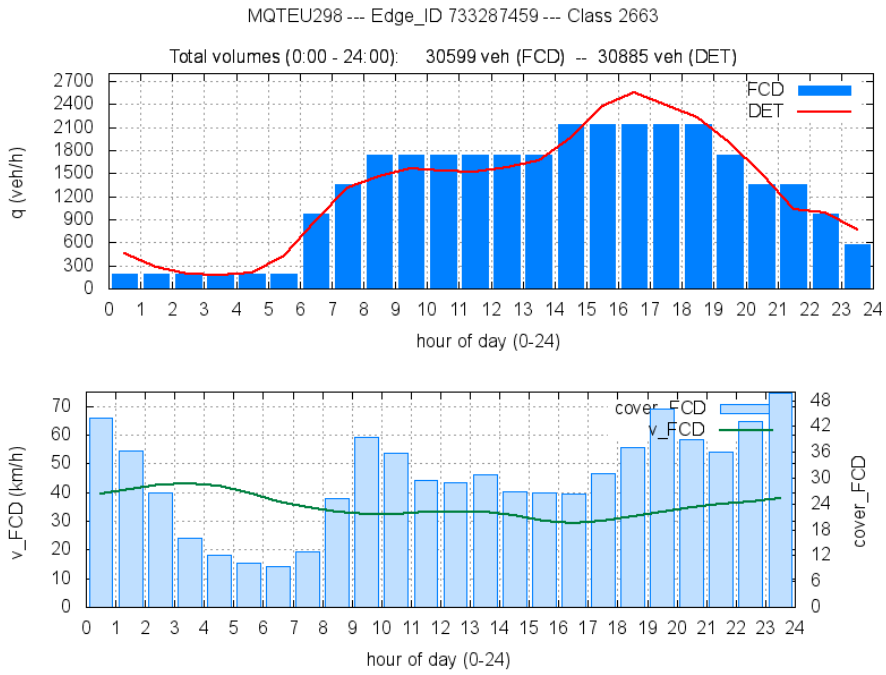


Figure 7: Comparison of measured flow with model results for good FCD coverage.

While roads with good FCD coverage and medium or high saturation level show good results (see Fig. 7), for roads that do not fit these criteria results were often not usable. Reasons for this are either insufficient coverage from FCD results or low changes in speed depending on the flow in under saturated conditions. In an extension of this approach Bayesian Networks were used to model traffic states and temporal dependencies in the transition between these states [Neu13a] which provides even more reasonable results for traffic flows. Further research will take the effect of variable message signs and traffic signals into account and will optimize the road classification schema used so far.

4 Summary

While basic algorithms for FCD processing usually have only a limited scope which is focused on one vehicle and a limited amount of positional data, FCD of higher quality or further information can be gained by extending this scope. This extension could be either temporal by inclusion of historic data or spatial by using whole trajectories instead of small parts, to determine turn relations or construct routes as new ground truth for quality evaluations. Data collected by new devices will improve the amount and availability of FCD as well as improve the quality of raw data by including additional information and through higher sampling rates. Changes in underlying network models can be used to generate more fine granular results by assigning travel or delay times more accurately. Based on these data new applications can be created (e.g. use of FCD for operational traffic management or to determine driver acceleration profiles) or existing ones improved due to better quality of traffic data generated by FCD.

References

[Bro07] E. BROCKFELD, W. WAGNER and B. PASSFELD: "Validating Travel Times calculated on the basis of taxi floating car data with test drives". In: *14th ITS World Congress*. Beijing, China, Oct. 9–13, 2007.

[Bro10] E. BROCKFELD, T. NEUMANN, A. SOHR, and G. KUHNS: "Turn Specific vs. Link Based Travel Times calculated from Floating Car Data". In: *12th World Conference on Transportation Research*. Lisbon, Portugal, July 10–15, 2010.

[Hen07] D. A. HENSHER and K. J. BUTTON: *Handbook of Transport Modelling*. 2nd Edition. Emerald, Inc, 2007.

[Her09] J. C. HERRERA, D. B. WORK, R. HERRING, X. BAN, Q. JACOBSON, and A. M. BAYEN: "Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment". In: *Transportation Research Part C* 18 (2010), pp. 568–583. URL: http://dx.doi.org/10.1016/j.trc.2009.10.006

[Kuh11] G. KUHNS, R. EBENDT, P. WAGNER, A. SOHR, and E. BROCKFELD: "Self Evaluation of Floating Car Data Based on Travel Times from Actual Vehicle Trajectories". In: *IEEE Forum on Integrated and Sustainable Transportation Systems*. Vienna, Austria, June 29–July 1, 2011.

[Lom76] N. R. LOMB: "Least-squares frequency analysis of unequally spaced data". In: *Astrophysics and Space Science* 39 (Feb. 1976), p. 447–462.

[Neu13a] T. NEUMANN, P. BÖHNKE, and L. TOUKO: "Dynamic representation of the fundamental diagram via Bayesian networks for estimating traffic flows from probe vehicle data". In: *16th International IEEE Annual Conference on Intelligent Transportation Systems*. The Hague, Netherlands, Oct. 6–8, 2013.

[Neu13b] T. NEUMANN, L. TOUKO, P. BÖHNKE, E. BROCKFELD, and X. BE: "Deriving Traffic Volumes from Probe Vehicle Data using a Fundamental Diagram Approach". In: *13th World Conference on Transportation Research*. Rio de Janeiro, Brazil, July 15–18, 2013.

[Pre07] W. H. PRESS, S. A. TEUKOLSKY, W. T. VETTERLING, and B. P. FLANNERY: *Numerical Recipes.* 3rd Edition. Cambridge: University Press, 2007.

[Soh09] A. SOHR, P. WAGNER and E. BROCKFELD: "Floating Car Data based travel time prediction with Lomb periodogram". In: *Proceedings 16th World Congress on ITS.* Stockholm, Sweden, Sep. 21–25, 2009.

[VAe95] M. VAN AERDE: "A Single Regime Speed-Flow-Density Relationship for Freeways and Arterials". In: *74th TRB Annual Conference*. Washington, D.C., USA, 1995, paper ID 950802.

[Vor06] P. VORTISCH: "Modellunterstützte Messwertpropagierung zur Verkehrslage-schätzung in Stadtstraßennetzen". In: *Schriftenreihe des Instituts für Verkehrswesen der Universität Karlsruhe*. Vol. 64. Karlsruhe, 2006.

Corresponding author: Günter Kuhns, German Aerospace Center (DLR), Institute of Transportation Systems, 12489 Berlin, Germany, phone: +49 30 67055 216, e-mail: guenter.kuhns@dlr.de

The Impact of Loop Detector Distance and Floating Car Data Penetration Rate on Queue Tail Warning

Gerdien Klunder[1, 2], Henk Taale[1, 3], Serge Hoogendoorn[1]

[1] Technical University of Delft

[2] TNO

[3] Ministry of Infrastructure and the Environment, TrafficQuest

Abstract

There is a growing interest in the traffic community about the relation between traffic data quality and the efficiency of traffic management. Data collection is expensive and if the same level of traffic performance can be reached with less data or if traffic management becomes more efficient with better data, then that is interesting for a lot of transport organisations. In this paper the problem is introduced and illustrated by presenting the results of a study into the effect of different loop detector distances and floating car data (FCD) penetration rates on a queue tail warning system. It shows that for a detector distance of more than 300 meters the performance deteriorates quickly and that the addition of only 1% FCD increases the performance considerably.

Keywords: Floating Car Data, Traffic Data Quality, Queue Tail Warning

1 Introduction

Generally speaking, more and more data is coming available. In a study from IBM [IBM11] it was stated that 90% of the data in the world of today has been created in the last two years alone. As a consequence, in just a few short years the challenge has shifted from 'if we only had the data' to 'how can we derive better intelligence from the data' (K. T. PARKER, President and CEO VIA Metropolitan Transit). The growth in data also holds in the traffic world. Not only more data is coming available, but also different types of data from different sources, such as loop detector data, floating car data (FCD), GPS or GSM data, blue tooth data etc. Especially, floating car data is a rapidly growing data source, fed by the recent growth of smartphones and smartphone GPS applications.

Dynamic traffic management and information is used by road operators in order to improve network utilization, safety or the environment. Examples are influencing the traffic

flow by influencing speeds, lane use, route choice or merging operations by employing variable message signs (VMS), Dynamic Route Information Panels (DRIPs), ramp metering etc. In order to operate the measures, to generate traffic information and to choose the best suitable measure, traffic data are required. Accurate, reliable, high quality traffic data are a prerequisite for effective traffic management and information services.

Each data type has its own characteristics and quality. The required quality for a dynamic traffic management (DTM) measure or traffic information service differs, depending on the type of measure or information needed. Some measures are more time critical than others, while also the required accuracy requirements may differ. However, good research to establish requirements for the quality of traffic data in relation to the intended traffic management goals is lacking, while more and more new traffic data is coming available and the demand for reliable traffic information is increasing. Therefore more research on this subject is needed.

If the requirements for traffic data can be determined accurately for certain traffic management applications, this will give new possibilities for better traffic management: It will lead to a better achievement of the traffic management goals with the same data, i.e. more efficient data use. Also, better requirements for data acquisition can be imposed to traffic data providers, which may lead to cost reduction when less detailed/accurate data is sufficient, or when data acquisition can be tuned better for better results. For example by choosing optimal monitoring locations. An advanced possibility to improve the performance of traffic management applications is to select dynamically the best algorithm and data processing technique for the current situation and available data.

In this paper some background on the topic is given and some developments are described. After that the topic is illustrated with the relation between different spatial resolution data of loop detectors and floating car data on the performance of a queue tail warning system. The queue tail warning system was chosen, because it is a widely applied system in the Netherlands that uses dynamic speed limits on overhead matrix signs to warn drivers about downstream congestion. The system now operates on data from (induction) loop detectors, which have been installed widely on the Dutch motorway network. However, for cost saving reasons, one is interested if the system can function well enough with less loop detectors and/or with the use of other data sources. A first analysis of this interesting case is presented in this paper.

2 Background

2.1 State-of-the-art

An important development concerning collecting and distribution of traffic data in the Netherlands is the National Data Warehouse for Traffic Information (NDW), in which road authorities work closely together to develop and exploit a database for traffic data and to effectively use this data for traffic management and traffic information. The NDW collects, processes, stores and distributes all relevant traffic data to provide complete, reliable and up-to-the-minute information on the status of the main Dutch road network. Quality

requirements have been defined by the NDW and imposed to traffic data suppliers. Currently, there are discussions about redefining the quality requirements, especially to differentiate them for different road types or traffic management applications, because the current quality requirements cannot always be met and will lead to high costs, as presented in [Fel12].

In [Klu12], a preliminary study was performed on the relation between inaccurate traffic data and route choice, which concluded that accurate traffic counts are important for route choice information in case both route alternatives are close to oversaturation. In [Tam11], a study was performed on the relation between data quality and dynamic traffic management. However, this research studied only the effect on the resulting information or traffic management measure, not the impact on the traffic system, and they concluded that more thorough research is needed on this. Also at European level it has been identified that there is a lack of common quality criteria for traffic data and services. The QUANTIS project [Öör10] aimed to provide preliminary insights into the issue. Also in the U.S. it is recognized that the matter of data quality has become more urgent in recent years by the increase of ITS applications and various travel information systems, as reported in the “Data Quality White Paper" from the Federal Highway Administration [Ahn08].

Concerning the use and comparison of induction loop data and FCD data, research had been done already for example in [Gaz71]. In this article, a new method is put forward for fusing heterogeneous and semantically different data from different traffic sensors. In [Lin07] they compared and used both induction loop data and FCD for traffic state estimation, and also performed a cost-benefit estimation.

2.2 Organizational aspects of data monitoring

Apart from the quantitative aspects, also organizational aspects are important, because many different parties need to cooperate in order to get access to the different data sources, to define data format standards and to implement data processing algorithms. These include private parties who collect traffic data, such as navigation system providers, and public parties like road operators and traffic management centres. It seems that while data fusion techniques have been developed since the seventies of the previous century [Lin09], still few of them have been implemented in practice. Probably the cause of this is both a lack of good data as well as organizational problems.

Furthermore, the current operational traffic management systems such as the queue tail warning system, have been developed many years ago and in the meantime the systems and algorithms have evolved to such a complexity that it is not easy to switch to another (more efficient) system. If the current situation would be totally blank without any monitoring system, one could design a much more efficient traffic management system then the current one. In order to make this switch now, high initial costs are needed and many organizational issues will need to be solved. As such, the Netherlands has to deal with the law of the handicap of a head start, being one of the countries with the most extensive and oldest traffic monitoring system. In that sense, countries that don’t have an extensive monitoring system yet have an advance to design new efficient traffic management systems combining old and new data sources.

3 Application for queue tail warning

On the Dutch motorway network a queue tail warning system is applied (also known as AID, Automatic Incident Detection), which has the aim to prevent (secondary) accidents at the tail of traffic jams by lowering the maximum speed for vehicles approaching the traffic jam. A side benefit is that it helps to solve congestion quicker, especially shockwaves, because it reduces the inflow to the queue. It does this by detecting a traffic jam (low speeds), and gradually lowering the maximum speed upstream of the traffic jam tail. The first sign where the traffic jam is measured shows 50 as maximum speed, the next sign upstream 50 with flashers and the next sign upstream 70 with flashers. The portals are placed at a distance of around 500 meters from each other. It uses the available loop detection monitoring system as input and portals with variable message signs that show the maximum speed to the drivers. The system is already operational since the seventies of the previous century and proved to have lowered the number of head-tail accidents due to traffic jams. Based on research in 1984 [Bos07], the number of accidents was lowered with 16% in total, 36% of secondary accidents and 19% less vehicles involved in accidents.

The algorithm is based on speed detection of individual passing vehicles. First, outliers are filtered (speeds higher than 200 km/h are removed and speed slower than 18 km/h are set to 18 km/h). The algorithm works on reversed speeds instead of speeds, because that responds faster to speed differences for small speeds [Kli11]. A weighted moving average is calculated of the reversed speeds to smooth out speed fluctuations, by weighting the current smoothed speed with the current measured speed with a certain factor. This factor is higher for the measured speed when the new measured speed is smaller than the smoothed average speed then when the new measured speed is larger. In this way the system responds faster to low measured speeds than to high measured speeds. The system is triggered to start when the smoothed average speed gets below 35 km/h on one of the lanes, based on at least n vehicles. In the current research, $n=3$ is chosen. The trigger to turn off is when the average speed on all lanes gets above 50 km/h. This last condition is chosen in order to prevent too frequent on-off behavior of the system. The algorithm is responsive and not predictive: it is activated after the congestion has arisen and turned off after the congestion has been solved.

Though the system has proven to be successful, it is complex and expensive for maintenance. It needs a high density loop detection monitoring system which is currently under investigation in the Netherlands for lower cost alternatives, as explained before. Also, in other countries there usually is a much less dense monitoring network available. This justifies the current research into the performance of the system for different detector densities and including other data sources such as FCD.

3.1 Research questions for the queue tail warning case

In order to determine the effect of different spatial resolutions of detector data and penetration rate of FCD data on the performance of the queue tail warning system, calculations have been done with a detailed real-world dataset. Details of this dataset are given in the next paragraph. It was used to answer several main questions:

- Which loop detector distance is needed for a sufficient performance of the queue tail warning system?
- With which penetration rate of FCD is it possible to reach a comparable performance?
- Which improvement is possible for a combination of FCD and loop detectors?

3.2 Real-world data

As a test dataset, empirical microscopic loop data from a densely used motorway in the U.K. are used. The data come from the Active Traffic Management section of the M42 motorway near Birmingham [Wil11]. This section has an unprecedented coverage of inductance loop detectors, with a nominal spacing of around 100 m. During 2008/09, 16 consecutive detectors on the northbound carriageway were enhanced so that, among other improvements, the full individual vehicle data of all vehicles driving through the 1-mile section were recorded. For the research described in this paper a dataset of 10 days (1st to 10th October 2008) was used for a motorway stretch of one kilometre which contained 10 detectors.

The individual vehicle data include the passage time, speed, lane number, and length of each vehicle as it passes each of the sites. With this high resolution, one can track most individual vehicles through the section in most traffic conditions and thus in effect reconstruct their trajectories [Wil08]. As such, a floating car data set was constructed by interpolating the individual vehicle recordings between the detectors. The FCD data was subsequently generated by sampling the trajectories at a resolution of one Herz. During the 10 measured days, a sufficient amount of congestion and shockwaves occurred to test the AID algorithm.

3.3 Experimental set-up

Since the goal of the queue tail warning algorithm (AID) is to prevent accidents when approaching the tail of the traffic jam, the performance of the system should ideally be tested in practice by counting the number of accidents over a long period of time. Since this is a long and unreliable process and doesn't allow for experimenting, the performance is checked using indicators that are related to the safety of the vehicles approaching a traffic jam. These are the time to detection of the traffic jam, the error in the estimated location of the tail of the traffic jam and the number of detected traffic jams. In this study time to detection is defined as the difference between the first time of detection of the traffic jam (average speed < 35 km/h) in the baseline situation and the situation under investigation, where the baseline scenario is defined as the 100% FCD scenario, since this provides complete information about the traffic state. The error in the estimated location of the tail of the traffic jam is defined as the difference between the most upstream location of the detected traffic jam in the baseline and the situation under investigation at the same time. The number of detected traffic jams is defined as the number of times that the AID algorithm was triggered to go on (ones it is on, it needs to go off before it can be triggered to go on again). The idea behind

these indicators is that the risk reduction is larger when there are less vehicles that approach a traffic jam without passing the lower speed warning of the system (or equivalently, when there are more vehicles warned by a lower speed limit).

To test the effect of the detector distance on the performance, several distances have been tested by leaving out the detector data of part of the detectors. Since the length of the measured motorway stretch is 1 kilometre and contains 10 detectors, only a limited number of detector configurations were possible. The following detector distances have been used: 1000 m, 550 m, 385 m, 288 m, 192 m and 97 m. This is done by selecting a subset of the detectors (2,3,4,…,10 detectors) with as much as possible equal distances in between, covering as much as possible the full length of the measured section. For example, for 1000 m, the first and last detector have been selected, for 550 m the middle detector was selected, for 385 m the fourth and eighth detectors were selected, and so on.

Since the basic AID algorithm has been developed for loop detector data, it is as such only suitable for data measured at fixed locations. In order to be able to apply it with FCD data, some additions were needed to the algorithm. This has been done as follows: the FCD second-by-second data was interpolated at fixed locations, namely at every meter. The AID algorithm was now applied at each meter (as if there was a detector at every meter). Again at least three vehicle measurements are needed to trigger the system. In this way, the location of a vehicle driving with low speed can be detected very accurately, though with low penetration rates the time to detection of a queue could be long.

The penetration rate was varied by a random draw (uniform, one draw per penetration rate) of all measured vehicles and taking into account only the data of this selected set of vehicles. The following FCD penetration rates have been simulated: 0%, 1%, 2%, 10%, 50% and 100%. Also combinations of FCD and loop detector data have been simulated. This was easily possible in the above explained algorithm, by applying the algorithm both for all vehicle measurements at the loop detector locations and for the set of FCD vehicles at every meter.

By using 100% FCD, the exact moment of all congestion occurrences and locations of the traffic jam tail have been determined. To determine the ground truth, the location and timing of commencement of the traffic jam tail was determined at every second and every meter as the most upstream location where the AID was triggered on.

4 Results

The results of the analysis are shown in Figure 1, 2 and 3. Looking at the detection rate in Figure 1, a 100% penetration rate logically shows a detection rate of 100%, while loop detectors without FCD only detect up to 30% of the congestion in the base case. This large difference is caused by the high resolution (1 second and 1 metre) of ground-truth congestion and as a result the on-off behaviour with the high-resolution FCD. A penetration rate of 50% FCD detects 60%-75% of the traffic jams.

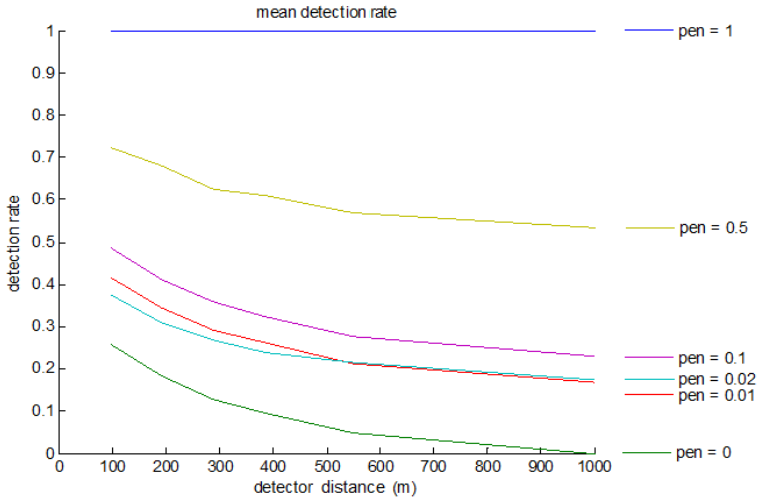


Figure 1: Detection rate for various detector distances and penetration rates of FCD vehicles.

As shown in Figure 2, the time to detection varies from 10 seconds to 100 seconds without FCD, while with 50% FCD the detection time stays below 40 seconds. Also the location error benefits from FCD data, as is shown in figure 3. While with loop detectors the location error increases up to 250 meters, with the addition of 1% FCD this is reduced to 200 meters, and with 50% FCD it stays below 80 meters.

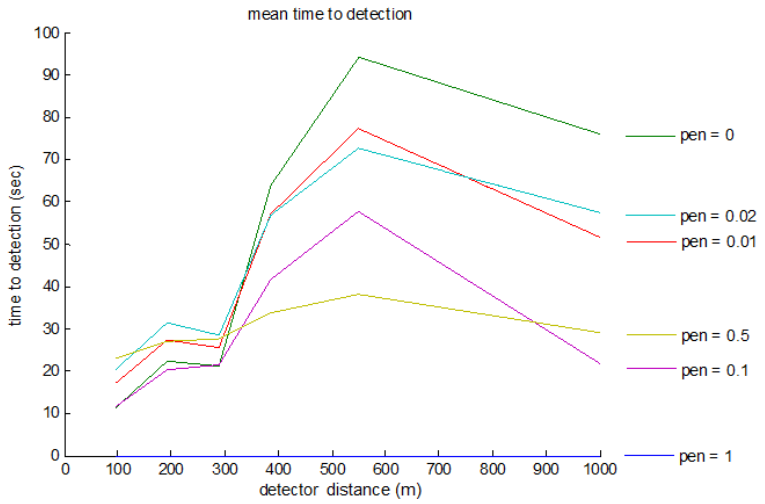


Figure 2: Time to detection for various distances and penetration rates.

It may seem strange that the time to detection goes down after 550 meters. This is probably a boundary effect because two detectors where used (one at the upstream boundary and one at the downstream boundary) which capture traffic jams better than one detector in the middle for the case of a detector distance of 550 meters. Another remark can be made about the influence of flow on the results. In the used dataset, the flow was rather high; this is to be expected in a situation where shockwaves occur. However, in some cases (such as when an incident occurs) the flow can be much lower while still congestion will form. Using low penetration FCD for Queue Tail Warning will be less effective in this case, since the probability will be larger that the minimum detection boundary of the algorithm (three FCD vehicles with a speed below 35 km/h) will not be achieved.

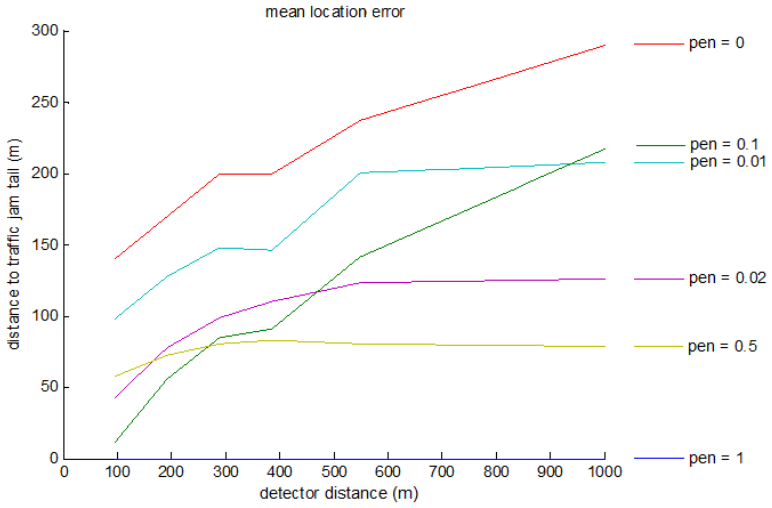


Figure 3: Location error for various distances and penetration rates.

5 Conclusions

Linking traffic data quality to the efficiency of traffic management is an unexplored field. While more and more traffic data are coming available, not much is known about the needed data quality in order to reach the desired goals of traffic management. If the requirements for traffic data can be determined accurately for certain traffic management applications, this will give new possibilities for better traffic management. It will lead to a better achievement of the traffic management goals with the same data, i.e. more efficient data use, and cost reduction, for example when less detailed/accurate data can be sufficient. However, in order to achieve this in the current world of traffic management practitioners, a change of view is needed: start with what you want to achieve, instead of what data you have.

Looking at the results of the data study to the effect of different loop detector distances and FCD penetration rates on a queue tail warning system, we can answer the research questions stated in paragraph 3.1 as follows:

The first question was which detector distance is needed for a sufficient performance of the queue tail warning system. Up to 300 meter detection distance, the performance seems to be reasonable: the detection time stays below 25 seconds and the location error below 200 meters. With larger detector distances, the time to detection and location error increase quickly.

The second question, with which penetration rate of FCD is it possible to reach a comparable performance, it can be concluded that the detection time and location error is already shorter with 1% FCD.

Thirdly, which improvement is possible for a combination of FCD and loop detectors? Looking at a detector distance of 500 meters, adding 1% FCD reduces both the detection time and the location error with 20%.

It must be noticed though that the used indicators are related to the final aim, i.e. increasing traffic safety, but the exact relationship is not known.

6 Further research

Further study is needed to determine the relation between the used indicators and the effect on traffic safety, i.e. the relation between the time to detection and location of the traffic jam tail in combination with reduced speed limits on the risk of traffic jam tail collisions. Options to study this are for example driving simulator studies, camera observation in practice or using surrogate safety measures in a traffic micro simulation study.

Also more accurate results could be achieved with a larger dataset and more experiments to reduce stochastic effects. The presented results are based on data from a quite short road section and also influenced by the random draw of FCD vehicles. The long computational time of the experiment (due to the very detailed data of individual trajectories on a 1 Hz base) made it too time consuming to repeat the experiment for a high number of random draws. Furthermore it would be more realistic to use a larger set of real-world measured FCD on a longer track. Another approach to overcome the problem of a limited dataset would be to use simulation. However, simulation models need to be calibrated thoroughly with real-world data as well in order to be sufficiently representative.

This research is part of a PhD research, which aims to address the problem of the relation between traffic data quality and traffic management/information in a broad perspective. Therefore, in future research quality requirements will be established for several traffic management and information applications and situations. This will be done both for time critical applications such as ACC, medium time critical applications such as queue length estimation for urban control and less time critical applications such as routing and network-wide traffic management. In order to be able to generalize the results, a general framework will be designed. Also the type of errors that occur in reality on different types of traffic data will be investigated, as well as statistical relations between different types of errors.

References

[Ahn08] K. AHN, H. RAKHA, and D. HILL: *Data Quality White Paper*. Federal Highway Administration, 2008.

[Bos07] N. S. VAN DEN BOSCH: "Blik op het Wegennet. Mogelijkheden van videomonitoring in Nederland". Graduation report. Delft: Technische Universiteit Delft, 2007.

[Fel12] E. FELICI and O. VROOM: "Differentiation of data quality for Dynamic Traffic Management". In: *19th ITS World Congress*. Vienna, Austria, Oct. 22-26, 2012.

[Gaz71] D. GAZIS and C. KNAPP: "On-line estimation of traffic densities from time-series of flow and speed". In: *Transportation Science* 5 (1971), pp. 283–301.

[IBM13] IBM: *IBM big data platform*. (Last Access: 12 April 2013). URL: `http://www-01.ibm.com/software/data/bigdata/`.

[Kli11] I. KLIJNHOUT: "Motorway Control and Signalling: The Test of Time". In: *Traffic Engineering and Control* 25.4 (1984). URL: `http://www.highways.gov.uk/knowledge/1334.aspx`

[Klu12] G. A. KLUNDER, S. P. HOOGENDOORN, and H. TAALE: "The relation between data quality and traffic management investigated for a simple route choice model". In: *LATSIS Symposium 2012, 1st European Symposium on Quantitative Methods in Transportation Systems.* Lausanne, 2012.

[Lin07] J. W. C. VAN LINT and S. P. HOOGENDOORN: "The technical and economic benefits of data fusion for real-time monitoring of freeway traffic." In: *World Congress of Transportation Research*. Berkely, CA, USA, June 2007.

[Lin09] J. W. C. VAN LINT and S. P. HOOGENDOORN: "A Robust and Efficient Method for Fusing Heterogeneous Data from Traffic Sensors on Freeways". In: *Computer-Aided Civil and Infrastructure Engineering* 24 (2009), pp. 1–17.

[Öör10] R. ÖÖRNI, S. INNAMAA, R. KULMALA, A. KELLERMANN, R. EBNER, and D. NEWTON: "Evaluation report for optimal data quality in selected European service cases". In: *Deliverable 6 of QUANTIS - Quality Assessment and Assurance methodology for Traffic data and Information Services.* Espoo, Finland, 2010.

[Tam11] C. TAMPÈRE, F. CORMAN, and W. HIMPE: "Invloed van datakwaliteit op Dynamisch Verkeersmanagement". Report. Leuven: Katholieke Universiteit Leuven, 2011.

[Wil08] R. E. WILSON: "From inductance loops to vehicle trajectories". In: *Proc. Symp. Fundam. Diagram – 75 years*, 2008, pp. 134–143.

[Wil11] R. E. WILSON: *Trafficdata: Individual Vehicle Data from the M42 Motorway*. (Last Access: 11 July 2011). URL: `http://www.enm.bris.ac.uk/trafficdata`.

Corresponding author: Gerdien Klunder, Delft University of Technology, 2628 CN Delft, The Netherlands, +31 15 27 85440, e-mail: g.a.klunder-1@tudelft.nl

Route Choice Identification and Selection from Sparse Floating Car Data Sets

Gennaro Ciccarelli[1], Claudia Castaldi[1], Chiara Colombaroni[2], Gaetano Fusco[1]

[1] University of Rome La Sapienza

[2] University of Rome Niccolò Cusano

Abstract

The paper deals with the opportunities and difficulties to exploit large sets of sparse floating car data for modeling purposes, more specifically for route choice analysis. A methodology is introduced for path identification and selection. It explores all possible routes between an origin-destination pair starting from a set of sparse observed vehicle positions; it identifies the most likely routes for each trip and finally selects a limited set of representative paths that appear significantly different on the road network model. Finally, an application is presented on a set of 62 routes between one origin-destination pair, selected from a database of several million of trips tracked in the metropolitan area of Rome. The corresponding set of representative paths is shown which provides the best balance of complexity and accuracy in representing users' behavior on the road network model.

Keywords: Floating Car Data, Route Choice, Path Set Generation, Representative Routes

1 Introduction

The increasing diffusion of vehicular and personal satellite positioning devices supplies a huge amount of floating car data, which provides an unprecedented detail of vehicular traffic on the road network and users' mobility patterns. The analysis of repeated observations of trips performed by many individuals can disclose many aspects of travellers' mobility behaviour, such as route choice process, as trip chaining propensity to a day-to-day revision of previous choices. These aspects have been until now difficult to observe directly on large samples of users or uncertain to estimate in their actual extent. Other than giving new insights on analysts' knowledge about mobility patterns and perhaps inspiring the development of new models, floating car data can be exploited to validate the numerous behavioural models that were developed in the last years. This is mainly true for route choice models, which are usually calibrated on small samples of users' route choices and validated on the basis of their aggregate results on link flows. However, repeated observations of current road performances and corresponding route choices by the same

user provide a direct source of information for a thorough calibration of the random utility models, which means solving the following problems: determine the most reasonable method for generation of the path choice set; specify the most suitable mathematical structure, which indeed should capture the correlation among different paths; determine the most likely values of the coefficients of the choice model. In order to reduce the number of paths that compose the choice set, various indicators have been proposed that measure the dissimilarity of route alternatives. [Akg00] introduced the dissimilarity in terms of length of shared links between two paths. [Del05] used the concept of a buffer zone to characterize heterogeneous paths. [Mar09] proposed an indicator evaluated at nodes, which eliminates problems related to the buffer area and is more representative of the drivers' choice behavior. The problem of route dissimilarity is closely related to the covariance analysis. [Cas96] were the first to capture the correlation between route alternatives explicitly. They introduced a correction attribute, called commonality factor, in the deterministic part of the logit model formulation, which is proportional to the overlap of each generic path with the other paths in the choice set. [Bek02] adapted a logit kernel model to the route choice problem; [Mar04] developed a link based path-multivel logit model. [Cas01] reduced the computational complexity of choice set generation by viewing the choice set as a fuzzy set in an implicit model of availability/perception of choice alternatives. The management of large data sets of floating car data gives rise to some computational problems that require pre-processing data. Floating car data are taken as successive geographical coordinates and have to be matched on the road network before being applied in transport modelling. Although many map matching algorithms have been developed in the last years for navigation systems, they are not suitable for statistical analyses. In fact, floating car data are usually collected with much lower frequencies than those applied by on-board navigation systems. Thus, the problem arises to recognize the route followed by the vehicle from sparse sample points. Specific methods for dealing with sparse positions data are to be implemented. Moreover, the most recent digital road maps developed for route guidance provide a so high level of detail that amplifies the problems due to the sparseness of position data without adding any useful information for analysis and modelling purposes. [Rah13] developed a two-step method that first applies a map matching algorithm to individuate a set of candidate links and then performs a path inference by connecting all matched points to build a candidate graph and finally finding the most likely path in such candidate graph. [Fre07] introduced the concept of subnetwork, which tries to capture the most important correlation among similar paths on the network without considerably increasing the model complexity. They assumed the choice set be composed by all possible paths on the network and developed a method for building the subnetwork by applying factor analysis.

In this paper, we focus on a quite different goal. Other than recognizing the most likely routes from sparse floating car data, we aim more specifically at identifying a limited number of significantly different paths that represent drivers' route choices with the level of accuracy required by traffic models. The paper is structured as follows. The next section explains the methodology applied for identification and selection of the route choice set. Related results are illustrated in Section 3. Final remarks summarize conclusions and provide suggestions for further development of the ongoing research.

2 Methodology for Route Choice Set Identification

The methodology developed to determine some significantly different paths that represent the actual route choice alternatives consists of the following operations:

- map matching of single "points" on the road network graph;
- path reconstruction, which explores a reasonable number of feasible paths connecting two successive projected points and identifies the most likely path for each trip in the data set;
- path selection, which analyzes the whole set of the reconstructed routes, splits it into several clusters and selects the most representative path for each cluster. Such representative paths compose the final route choice set of alternatives as it can be used for behavioral models.

The first operation has already been addressed in a previous project work using a semi-probabilistic map matching algorithm and illustrated in [Ram12]. The latter points are described in the following. The aim is to obtain a set of feasible paths representative of users' preferences and significantly different with each other.

2.1 Path Reconstruction

Data of vehicle trips have been stored in a database. Each trip is described by a sequence of records, which depict the instantaneous states of the vehicle and the travelled distance since the origin. Each pair of consecutive records belonging to the same trip forms a segment. For each segment, we calculate the *k*-shortest paths between the sampled points by implementing the algorithm designed by Russo and Vitetta [Rus06]. To identify the most likely path that represents the actual vehicle's route, we choose the path having the minimum difference of length with the observed travelled distance. Then, for each trip we reconstruct the whole route followed by the vehicle from the origin O to the destination D as the sequence of most likely paths from consecutive sample points. We also compute the difference between the measured distance of each segment and the corresponding value calculated on the graph.

The processing time is a critical issue in large databases of floating car data. The time for processing a single trip varies with the number of lines that compose it, that is, with the length of the trip and the level of detail of the graph. For the *k*-shortest paths calculation, we fixed the value k=7, after some experimental results, which showing that larger values of k increased the processing time considerably without producing a significant reduction of the error. We have executed these calculations sequentially, but an ongoing research consists of testing parallel algorithms that use General purpose Programming on Graphics Processing Unit (GPGPU), which is expected to reduce the computation time.

2.2 Path Selection

The path selection procedure takes in input the results of the path reconstruction routine, which supplies the set of routes most likely followed by the road users in their different trips; then, the problem is to select from the whole set of routes a subset of different paths that can be perceived by the users as different alternatives. This problem is solved by a heuristic algorithm that clusters the reconstructed routes in different sets and selects the most representative route on each set. The clustering criterion consists in maximizing the dissimilarity of paths belonging to different sets and minimizing that between paths of the same set. The following dissimilarity index $D(i, j)$ between route i and j is introduced

$$D(i,j)=1-\frac{1}{2}\left[\frac{L(P_i\cap P_j)}{L(P_i)}+\frac{L(P_i\cap P_j)}{L(P_j)}\right] \quad (1)$$

where $L(P_i)$ and $L(P_i\cap P_j)$ are the length of path i and the length of the overlapping part of paths i and j, respectively. Low index values indicate highly overlapping routes whilst unit values denote truly dissimilar routes. More complex indicators that introduce the travel time, the number and the category of links can be introduced. However, they require an extensive knowledge of the traffic speed on all links of the network in different hours of the day and take into account weekly and seasonal effects. However, such an in-depth knowledge of the road network performances is very difficult to achieve, so we prefer using the pure distance-based indicator.

After the final route choice set has been obtained, a representative route is chosen for each set. Such a route should represent at best drivers' choices and also the most relevant on the graph model; thus, a simple rule is applied that maximizes a weighted function of the users' frequency of choice and the relevance of the links travelled in the hierarchy of the road network model. The path selection procedure applies the following steps:

- Step 0 (Initialization). Get the set $\mathbf{P} = \{P_i;\ i=1, 2, \ldots, n\}$ of reconstructed paths and take it as the current set of representative paths, $\mathbf{S} = \mathbf{P}$. Initialize the number of path sets $m=1$. Let M be the desired number of path choice sets.

- Step 1 (Dissimilarity analysis). For each pair of paths P_i and P_j of S, identify the road links shared by P_i and P_j, and compute their dissimilarity index $D(i, j)$ from Eqn (1).

- Step 2 (First split). Find the most dissimilar pair of paths in the set S, say $(h,h')=\arg\max_{i,j}\{D(i,j)\}, i, j=1,2,\ldots,n$

  Take each of these two paths as the first item of two new distinct sets of paths, $S_1=\{P_h\}$ and $S_2=\{P_{h'}\}$. Update the number of sets $m=m+1$. Let $k=1,2,\ldots n_k$, with k=1,2...,m, be the cardinality of the set S_k.

- Step 3 (Path classification). For each path P_i, $i=1,\ldots, n$, find the set $S_l \in \{S_1, S_2, \ldots S_m\}$ of minimum dissimilarity with P_i

$$l = \arg\min_{k}\left\{\min_{j_k}\left[D(i, j_k)\right]\right\}, i = 1,2,\ldots, n;\ j_k = 1,2,\ldots n_k;\ k = 1,2,\ldots, m$$

and put P_i in S_l. If $m<M$, go to Step 4. Otherwise, go to Step 5.

- Step 4 (New set identification). For each set of paths $S_k \in \{S_1, S_2, \ldots S_m\}$ find the path P_{q_k} of maximum dissimilarity within each set and select that with the maximum value among them

$$D_{q_k} = \arg\max_{i_k}\left\{\max_{j_k}\left[D(i_k, j_k)\right]\right\};\ (i_k, j_k) = 1,2,\ldots, n_k;\ k = 1,2,\ldots, m$$

$$p = \arg\max_{k}\left\{D_{q_k}\right\}, k = 1,2,\ldots, m$$

Define a new path set S_p and put P_{p_k} in S_p. Update the number of sets $m=m+1$ and Go to Step 3.

- Step 5 (Selection of representative routes). For each set of paths $S_k \in \{S_1, S_2, \ldots S_m\}$ find the path P_{r_k} that maximizes the following function

$$r_k = \arg\max_{k} \frac{\sum_{a \in P_k} w_a f_a l_a}{\sum_{a \in P_k} l_a}\left(\frac{l_k^m}{\sum_{a \in P_k} l_a - l_k^m}\right)^{\beta},\ k = 1,2,\ldots M$$

where f_a and l_a are, respectively, the frequency of choice and the length arc a, while w_a is a weight proportional to the hierarchy of the arc a on the network, l_k^m is the length of the shortest path in set S_k and β is a coefficient. Select the path P_{r_k} as representative route of the path set S_k. End.

3 Experimental Analysis of Individual Route Choices

This section presents the results obtained by applying the method for path identification and path selection on a large database, composed by about 100 million of records, provided by tracking about 100,000 vehicles for one month in the metropolitan area of Rome. An onboard unit tracks vehicle positions and speeds at a high frequency rate but records the following data only every 2 km: vehicle identifier, timestamp, geographic coordinates, instantaneous speed, distance traveled from the previously sampled point. Unlike an aggregate analysis of mobility patterns, conducted on the whole database, the individual analysis is limited to the drivers traveling on a single Origin-Destination. Specifically, the O-D relationship with the greatest number of trips and a distance larger than 8 km is investigated. Such a data sample is composed by 62 trips made by 20 different GPS equipped vehicles travelling along the selected O-D destination during the month of May 2010. A more extensive analysis on the whole dataset is undergoing. The reference network graph is formed by 274,000 nodes, 32,948 arcs and 1,331 centroid nodes. The highest level of detail is in the urban area, which

covers an area of about 300km^2 and contains the largest number of graph elements so that the sampling points, collected every 2 km, are very sparse with respect to the detail of the road network.

3.1 Path Reconstruction Results

The results of the path reconstruction algorithm applied to the sample of 62 trips made by 20 users show that the reconstructed paths are composed on average by 171 road links and have an average estimated length of 13.3 km; the average travelled distance measured by GPS equipments is 12.1 km, with an average error of about 1.2 km, corresponding to about 10%. Figure 1 depicts the correlation between the observed and the measured length of all trips and highlights how larger errors are due to few outliers. Since *k*-shortest path algorithm always considers the path of minimum distance and the algorithm overestimates the measured distance, it is reasonable to think that so large errors are due to inaccuracies of GPS positions.

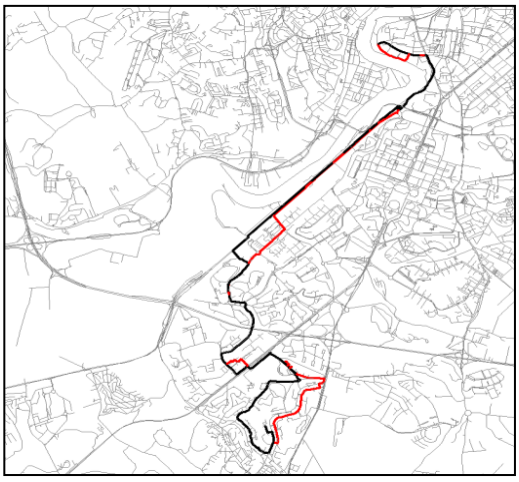


Figure 1: Maximum deviation between routes chosen by the same traveler

Display reconstructed paths on the road network highlights that many routes followed by the same vehicle differ by only slight differences. The two most dissimilar routes followed by the same vehicle depicted in Figure 2; they differ with each other by 30 road links and 1 km length. Their dissimilarity index is 0.45. The picture highlights also that the deviations between the two paths have not only a limited extent, but –more important for modeling purposes– they are restricted to local streets and concentrated on the initial and end extremes of the trip. It is unlikely that the driver perceived the two routes as two distinct alternatives, and it is likely that he or she decided slight changes to the path because of contingent factors, which are not worth being included in a behavioral model.

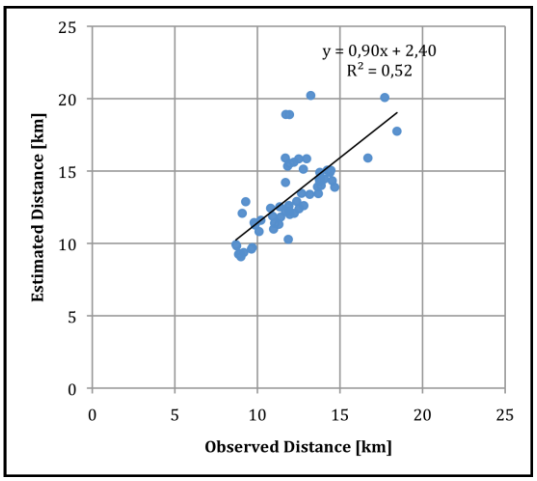


Figure 2: Correlation between observed and estimated travelled distance

3.2 Path Selection Results

To illustrate the performances of the path selection procedure in selecting the most representative routes to compose the path choice set, it is useful to compare the initial set of reconstructed routes (only a selection of 13 out of 62 is depicted in Figure 3) with the final set of 4 representative routes, shown in Figure 4 and obtained by assuming in the representativeness function β=1 and w_a =1 for all arcs a. The difficulty to recognize different relevant routes in the initial set explains the need for a systematic path selection procedure that clusters similar route alternatives and selects only the routes that represent significantly different alternatives for road users. The complex overlapping of reconstructed paths highlights the difficulty of the task.

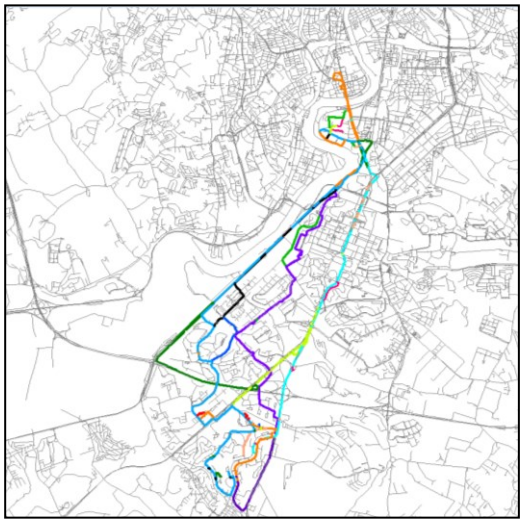


Figure 3: Visualization of (some) reconstructed routes on the road network

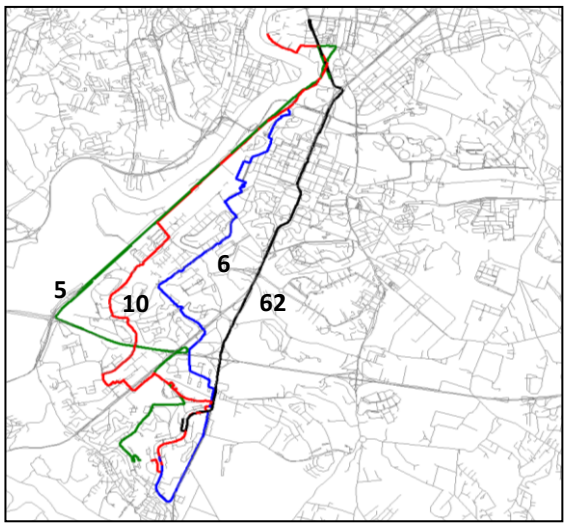


Figure 4 : Visualization of the final route choice set of representative paths

The four routes selected represent alternatives that are significantly different both spatially qualitatively. Quantitative characteristics are summarized in Table 1. Route 62 is the most direct and shortest alternative (11.3 km), which follows the main road artery in that urban sector, with many signalized intersections. Route 5 uses the local streets, and a stretch of the ring road expressway to follow a more distant radial artery with fewer intersections to go toward the city centre. Routes 6 and 10 have similar qualitative characteristics. They both are winding paths that use preferably minor roads and then join the less direct radial artery to reach the final destination. Frequency of choice indicates the relevance of each subset of routes in drivers' route choice. Route 10 represents a large number of similar routes (having a frequency of choice of 52%), routes 5 and 62 are chosen by about 24% of users. Route 6, although it has been chosen just once, is significantly different from the other ones and then has been selected as a relevant alternative. The last two indicators highlight the result of the clustering algorithm. The external dissimilarity is the average dissimilarity index of the paths of each subset with respect to the paths of the other subsets while the internal dissimilarity is the average dissimilarity index of the paths of each subset with respect to the other paths of the same subset. All the 4 subsets selected have an external dissimilarity higher than 0.85 (so, they are significantly different with each other) and an internal dissimilarity lower or equal to 0.45 (that is, they are homogenous, and their representative route stands for a large number of similar alternatives).

Table 1: Characteristics of the final route choice set

Route id	N. of choices	Rel. Frequency of choice	N. of Links	Length (km)	Avg. External Dissimilarity	Avg. Internal Dissimilarity
5	15	0.24	169	15.3	0.92	0.45
6	1	0.02	190	13.4	0.90	0.00
10	32	0.52	203	13.3	0.86	0.33
62	14	0.23	143	11.3	0.90	0.21

4 Final Remarks

The paper has illustrated a procedure for reconstructing vehicle paths from a large set of sparse floating car data and to select a limited number of significantly different routes that are representative of actual drivers route choices.

The research is still on-going and is continuing in the following directions: extend the analysis to the whole database; apply a parallel data processing for path identification; introduce a new dissimilarity index that takes into account road network hierarchy. Further research will be addressed to calibrate behavioural random utility models on observed patterns and develop dynamic process models to reproduce the day-to-day users' route choice behaviour.

References

[Akg00] V. AKGÜN, E. ERKUT, and R. BATTA: "On finding dissimilar paths". In: *European Journal of Operational Research* 121.2 (2000), pp. 232–246.

[Bek02] S. BEKHOR, M. BEN-AKIVA, and M. S. RAMMING: "Adaptation of logit kernel to route choice situation". In*: Transportation Research Record* 1805 (2002), pp. 78–85.

[Cas96] E. CASCETTA, A. NUZZOLO, F. RUSSO, and A. VITETTA: "A modified logit route choice model overcoming path overlapping problems: specification and some calibration results for interurban networks". In: *Proceedings of the Thirteenth International Symposium on Transportation and Traffic Theory*. Ed. by J. B. LESORT. Lyon, France: Pergamon, 1996, pp. 697–711.

[Cas01] E. CASCETTA and A. PAPOLA: "Random utility models with implicit availability/ perception of choice alternatives for the simulation of travel demand". In: *Transportation Research Part C: Emerging Technologies* 9.4 (2001), pp. 249–263.

[Del05] P. DELL'OLMO, M. GENTILI, and A. SCOZZARI: "On finding dissimilar Pareto-optimal paths". In: *European Journal of Operational Research* 162.1 (2005), pp. 70–82.

[Fre07] E. FREJINGER and M. BIERLAIRE: "Capturing correlation with subnetworks in route choice models". In: *Transportation Research Part B: Methodological* 41.3 (2007), pp. 363–378.

[Mar04] V. MARZANO and A. PAPOLA: "A Link based Path-multilevel Logit Model For Route Choice Which Allows Implicit Path Enumeration". In: *Proceedings of European Transport Conference.* Strasbourg, France, 2004.

[Mar09] R. MARTÍ, J. L. GONZÁLEZ-VELARDE, and A. DUARTE: "Heuristics for the bi-objective path dissimilarity problem". In: *Computers & Operations Research.* 36 (2009), pp. 2905–2912.

[Rah13] M. RAHMANI and H. N. KOUTSOPOULOS: "Path inference from sparse floating car data for urban networks". In: *Transportation Research Part C: Emerging Technologies*. 30 (2013), pp.41–54.

[Ram12] S. RAMBALDI, M. MARCHIONI, A. BAZZANI, and B. GIORGINI: "Traffic global analysis on the whole italian road network". In: *IEEE MIPRO, 2012 Proceedings of the 35th International Convention.* 2012, pp. 1678–1682.

[Rus06] F. RUSSO and A. VITETTA: *La ricerca di percorsi in una rete: algoritmi di minimo costo ed estensioni.* Milano: Franco Angeli, 2006. ISBN: 978-88-464-8228-0.

Corresponding author: Gennaro Ciccarelli, University of Rome "La Sapienza", Department of Civil Construction and Environmental Engineering, 00184 Rome, Italy, phone: +39 0644585145, e-mail: gennaro.ciccarelli@uniroma1.it

Optimizing Public Transport Planning and Operations using Automatic Vehicle Location Data: the Dutch Example

Niels van Oort[1, 2], Daniel Sparing[1], Ties Brands[2, 3], Rob M.P. Goverde[1]

[1] Delft University of Technology

[2] Goudappel Coffeng

[3] University of Twente

Abstract

There is a growing pressure on urban public transport companies and authorities to improve efficiency, stemming from reduced budgets, political expectations and competition between operators. In order to find inefficiencies, bottlenecks and potentials in the public transport service, it is useful to learn from recorded operational data. We first describe the state of publicly available transit data, with an emphasis on the Dutch situation. The value of insights from Automatic Vehicle Location data is demonstrated by examples. Finally, a software tool is described that makes quick comprehensive operational analysis possible for operators and public transport authorities, and was able to identify several bottlenecks when applied in practice.

Keywords: public transport, AVL data, service reliability, monitoring

1 Introduction

Similar to in other countries, public transport in the Netherlands has to face substantial cost-cutting measures. Although higher quality and more capacity are needed, funding for public transport is reduced. Improving public transport quality and extending capacity under reduced finances is a hard challenge, but we believe that several possibilities exist. The key factors to enhanced and more cost efficient public transport are travel time and service reliability.

Service reliability is the certainty of service aspects compared to the schedule as perceived by the user and is, next to travel time, one of the main quality aspects in public transport. (Potential) public transport passengers take these aspects explicitly in consideration while planning their trip mode [Oor13].

In several studies reliability-related attributes have been found among the most

important service attributes in a variety of situations. Balcombe et al. [Bal04] report that service reliability is considered by passengers twice as important as frequency. König and Axhausen [Kon02] conclude that the research done over the last decade shows that the reliability of the transportation system is a decisive factor in the choice behaviour of people.

Much research illustrates the potential of improving travel time and service reliability. One could think of improvements of vehicles, infrastructure, planning and operations. Literature shows that in urban public transport, substantial attention is given to ways to improve services at the operational level [Vuc05, Ced07, Oor09a]. Concerning strategic and tactical instruments, much research is already available on the planning instruments of priority at traffic lights, exclusive lanes and synchronization. The implementation of bus lane schemes and traffic signal priority are the most used solutions in this field (as shown by e.g. Waterson et al. [Wat03]). Both Ceder [Ced07] and Vuchic [Vuc05] present the different methods and effects, and also give an overview of the issues which need to be considered in synchronization. During the design of the schedule, optimising trip time determination and holding are potential instruments improving operational quality [Oor12, Del12, Xua11].

Less researched so far, but potential instruments are available during network design as well, for instance line length and design of terminals [Oor09b, Oor10]. A study of a new tram line in Utrecht, the Netherlands [Oor13], showed that about 65% of the (societal) benefits are related to these aspects, being over € 200 million during the total lifetime of the tram infrastructure. On busy bus trunks in Utrecht, a reduction of 30 seconds of trip time per bus saves about € 100,000–400,000 in operational costs per year.

The first step to increase operational performance is a proper analysis of historical operations. This paper focuses on bus and tram modes. Operations performance for heavy railways based on track occupation data is described in [Gov11]. Automatic vehicle location systems (AVL systems [Str00, Hic04]) are of great help to provide databases of historical performance with regard to travel time and reliability. Although such data has already been available for years to many operators, it is only recently that this valuable data is becoming available also to Dutch transit authorities, researchers and developers. In addition to facilitating analysis of performance, proper data also enables forecasts of future service quality [Kan08, Wil09]. In this paper we analyse a practical example of such a data set to illustrate the usefulness of these kinds of analyses: several bottlenecks are identified, providing transport authorities with insight into setting investment priorities.

2 Public transport and open data

Public transport companies have always dealt with large amounts of data when designing timetables, scheduling vehicles and staff, collecting fares and more recently tracking vehicle locations. However, it has only recently become possible to store large amounts of historic vehicle location and fare collection data, and therefore to analyse this data. Furthermore, in line with other “Open Data” initiatives in the public sectors, data related to public transport is currently becoming publicly available in more and more areas, notably in North America and more recently in certain European cities.

The first type of public transport data that became publicly available is timetable information. Besides supplying public transport route planners with timetable data, computer-readable timetable information also allows for efficient analysis and comparison of public transport networks, describing spatial coverage, commercial speeds, frequencies, and connections to adjacent public transport networks. Timetable information provides no insight yet, however, in the performance of the timetable realization and hence the service reliability of public transport and the real-time timetable information.

Accurate real-time vehicle location data (Automatic Vehicle Location systems, AVL) has become available for public transport operators with the wide availability of GPS and GSM devices. AVL data has also become publicly available in many areas in the recent years, albeit often with the condition that it is only used for passenger information. Early examples include the transit agencies of Washington, Boston and some other US bus companies. We note that these days most Western public transport operators provide some kind of real-time vehicle location (or expected vehicle arrival time) information to the public, but often this information is still not technically or legally available for storing or further processing by third parties.

2.1 The Dutch example

In the Netherlands, most public transport operators are on board with the initiative called Borderless Public Transport Information ("Grenzeloze Openbaar Vervoer Informatie", GOVI [Gov13]), aiming at making a wide range of public transport information available and processable from timetables to fares, vehicle location and punctuality. The data exchange interfaces ("*koppelvlakken*") are defined by the set of standards called BISON [Bis13]. Another source of open public transport information, such as a GTFS feed on the national level, is the company 9292 REISinformatiegroep BV [Rei13], a company specialized in passenger information owned by Dutch operators.

GOVI was designed to facilitate data communication between vehicles and the land side to enable dynamic passenger information. As an additional benefit, the actual and scheduled vehicle positions and times are logged in a database. Although this database was not the objective of the GOVI system, it is extremely helpful to monitor and analyze public transport performance.

In particular, in 2012, the first Dutch public transport operator agreed to legally release AVL data via GOVI for storage and analysis by third parties, such as researchers and developers [Gvb13]. Since then several other operators joined. Such data streams are in practice publicly available via the Dutch OpenGeo Foundation [Ope13]. The source of the data later presented in this paper is either directly from the transit authorities or via OpenGeo.

3 Insights from AVL data

As a first step, it is important to understand the structure and the quality of the data source. In our case, AVL data was available for several months from multiple operators in the format

described by interface KV6 of the BISON standard mentioned earlier. An example extract of the most important data attributes and the first and last few records related to a single public transport vehicle trip is presented in Table 1.

Table 1: Example data output from BISON interface KV6.

Time	Message type	Operator	Line	Journey	Stop	Punctuality
08:29:00	INIT	...	B120	7001	99990140	
08:29:00	ONSTOP	...	B120	7001	99990140	60
08:29:22	DEPARTURE	...	B120	7001	99990140	82
08:31:28	DEPARTURE	...	B120	7001	99990290	88
...						
08:51:04	ONROUTE	...	B120	7001		
08:52:37	ARRIVAL	...	B120	7001	99990500	-202
08:52:37	END	...	B120	7001	99990500	

This data table consists of timestamped messages of important events of the vehicle trip. In particular, a trip starts with an INIT initial message and ends with an END message, and all departures are logged with a DEPARTURE message. In case of some stops an ARRIVAL message is recorded too, allowing for an estimate of the dwell time. Furthermore, in case that there is no departure and arrival event taking place for a longer time duration (about a minute), an ONSTOP or an ONROUTE message is recorded, including exact location. Our data source already includes a value for delay, which equals to the difference of the message timestamp and the planned arrival or departure time.

Line-based analysis

A commonly used visualization [Fur06, Oor09a] of the performance of a transit line is plotting each trip as a line chart in a coordinate system of stops versus delay. Figure 1 (left) shows one month of bus trips of a certain line, as well as the median and 15th and 85th percentiles. Such a chart is useful to see both the level of variations in the execution of the timetable and the systematic deviations. Other phenomena that are shown by this particular chart are the ample time reserve used just before the last stop and the use of some holding points during the trips.

Another way to look at the same data is to plot vehicle headways instead of delays. A high frequency line with a high level of delays but regular headways remains attractive to the passengers. The chart is shown on Figure 1 (right), the scheduled headway is 10 minutes. This location-headway chart points out the regularity of high-frequency services along the line, as well as possible bus bunching.

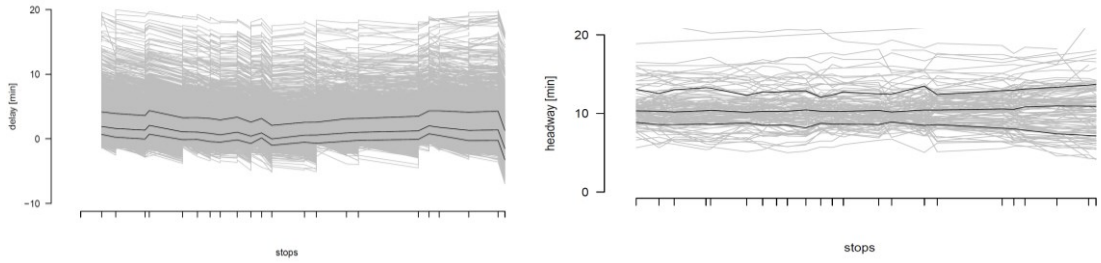


Figure 1: Vehicle delays (left) and headways (right) along a single route.

Network-wide analysis

The ubiquitous availability of vehicle locator devices allows one to take a step further from line-based performance evaluation and investigate patterns at the network level. Phenomena only visible on the network-level are the reliability of transfers, area-related issues and possible bunching or interference on multiple lines with shared sections. An example of a network-level data visualization is the average delay at each stop, including stops with several transit lines, shown on Figure 2.

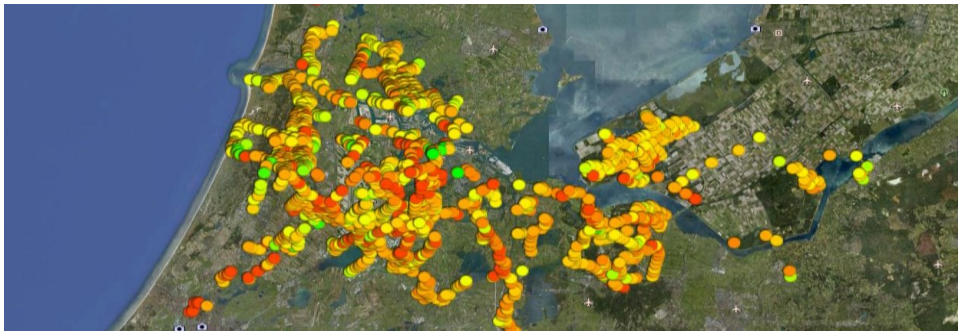


Figure 2: Average delay per stop (green: early, yellow: on time, red: late).

Inter-operator transfers

An aspect of public transport travel that was previously invisible to the public and to each operator, but of substantial importance to the passenger, is the reliability of transfers between multiple operators, such as between a long-distance train and a local bus. With open data, it is possible for anyone (so also to any operator) to investigate the actual reliability of inter-operator transfers – and for the operator to take steps if necessary.

Figure 3 shows a discrepancy between vehicle punctuality and passenger experience, for an example transfer that is scheduled to take 5 minutes excluding walking time. It is common that a public transport timetable includes a substantial time reserve before an important stop, and therefore as the left part of the figure shows, the vehicles are consistently early at the transfer stop. However, this means that the passengers structurally have to wait much longer at the transfer stop than they can expect from the timetable. As waiting time on the platform is perceived much less comfortable than in-vehicle time [Waa88], this means that trips including this transfer are perceived of a less quality than expected from the timetable.

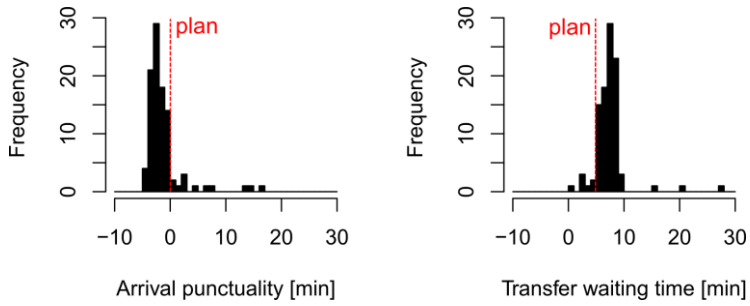


Figure 3: Arrival punctuality of a vehicle and transfer waiting time for the passenger at a transfer location.

The relevance of open AVL data with regard to improving transfers is the following: open information on the reliability of inter-operator transfers makes it possible for any operator and the transport authorities to gain insight into the reliability of these transfers and take steps if necessary, such as synchronizing timetables, holding vehicles in case of minor delays and informing passengers. See Sparing and Goverde [Spa13] for example for identifying transfers of interest and choosing which vehicles to hold in a multi-operator setting.

4 The GOVI-tool

To generate helpful insights from AVL data, the transport planning consultant firm Goudappel Coffeng developed a tool that translates data to information: the GOVI-tool. The GOVI database consists of all actual and scheduled arrival and departure times at all stops of all trips of the participating public transport operators in the Netherlands. This implies big data sets: a month contains 100.000-200.000 records per line. By subtracting the actual departure and actual arrival time, dwell time may be calculated. Comparing actual and scheduled times provide insights in the level of punctuality of the service. Finally, trip times may be calculated using departure times at a certain stop and arrival times at the following stop. Since information about stop distances is available as well, operating speed may also be generated.

To gain insight in the performance, mean values, 15- and 85-percentile values are calculated for the above mentioned aspects. This way, information about the variability is provided. The tool also provides information about cumulative values, such as total trip time, thereby illustrating the quality of actual performance compared to the schedule along the line. All information may be presented for every stop, per line and direction, period of the week (working day, Saturday or Sunday) and period of the day (AM peak, PM peak, daytime or evening).

In addition to showing the data in tables and figures, the tool is also capable of finding bottlenecks. The tool easily finds its way through all the data and selects (predefined) outlier values. The tool could for instance present a list of all stops where the average dwell time is larger than 30 seconds or where the schedule deviation is below zero (i.e. operating ahead of time).

The above described GOVI-tool has been used in the Utrecht region to analyze all bus lines, gaining insights in the actual performance and the largest bottlenecks. For 4 periods

per working day and Saturday and Sunday, insights where generated with regard to driving, dwelling and punctuality. Several performance indicators were compared to the Sunday values to gain insights into the maximum improvement potential. Below, we present examples of the generated graphs from both Utrecht and another line in North Holland, applying the GOVI tool (Figure 4-Figure 6): punctuality along a line, travel speed between stops and dwell times.

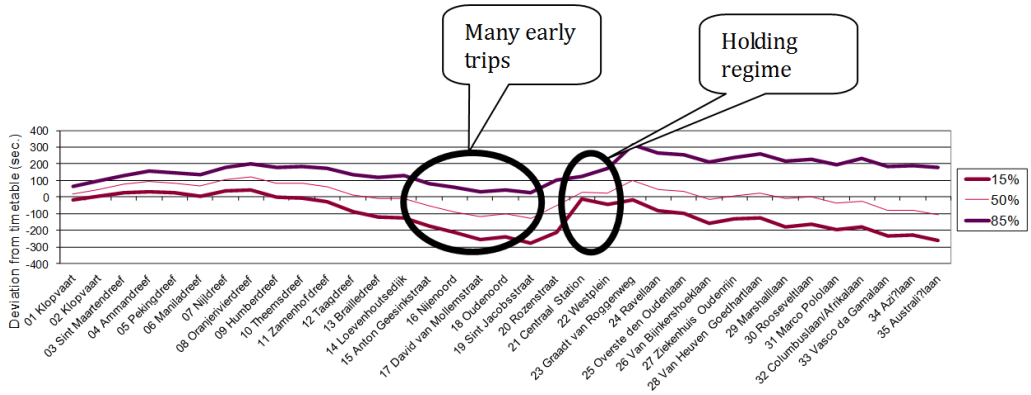


Figure 4: GOVI-tool graph of schedule deviation, bus line 7 Utrecht, evening rush hour.

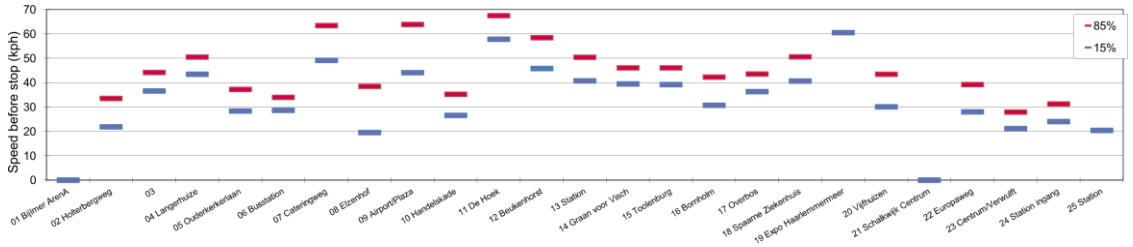


Figure 5: GOVI-tool graph of speeds between stops.

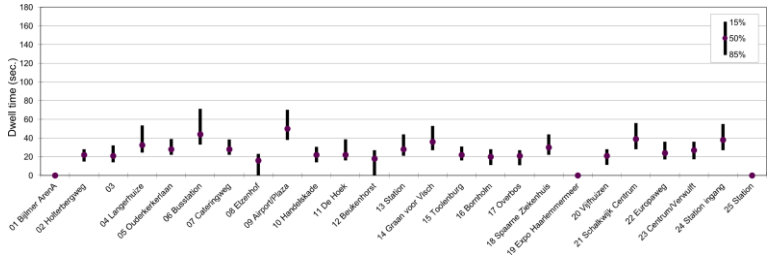


Figure 6: GOVI-tool graph of dwell times at stops.

5 Conclusions

While the funding of public transport is under pressure, the need to enhance quality is increasing. The key element to better and more efficient public transport are shorter and more reliable trip times. By removing bottlenecks of the operations, costs may be reduced while the quality will be increased, thereby increasing ridership and revenues. This way, the cost effectiveness of public transport is improved in two parallel ways.

To find the bottlenecks and the potential benefits of improvements, data of historical operations is required. In the Netherlands, this data is has become recently available via GOVI. The objective of the GOVI system was to facilitate dynamic travel information, but the recorded data also provide huge insights into actual and scheduled performance. Goudappel Coffeng developed a tool to translate all these data into information, so that bottlenecks can be identified and measure can be taken by the transport authorities to solve them, enhancing public transport.

A next step in the development of the tool is translating vehicle data into passenger impact. When additional data on passenger behaviour and flows is available, service reliability impacts per passenger per stop may be calculated (i.e. additional travel time and its distribution [Oor13]).

This research is performed in cooperation with BRU, the transit authority in the region Utrecht, the Netherlands; Delft University of Technology, Department of Transport & Planning; Goudappel Coffeng; the Netherlands Organisation for Scientific Research (NWO); and the Dutch OpenGeo Foundation. The authors thank their partners for their support.

References

[Bal04] R. Balcombe, R. Mackett, N. Paulley, J. Preston, J. Shires, H. Titheridge, M. Wardman, and P. White: *The demand for public transport: a practical guide*. 2004.

[Bis13] *Bison.* (Last Access: 18 July 2013). URL: http://bison.connekt.nl/.

[Ced07] A. Ceder: *Public transit planning and operation, theory, modelling and practice.* Haifa: Technion-Israel Institute of Technology, 2007.

[Cha03] J. Chang, J. Collura, F. Dion, and H. Rakha: "Evaluation of service reliability impacts of traffic signal priority strategies for bus transit". In: *Transportation Research Record* 1841 (2003), pp. 23–31.

[Del12] F. Delgado, J. C. Munoz, and R. Giesen: "How much can holding and/or limiting boarding improve transit performance?". In: *Transportation Research Part B: Methodological* 46.9 (2012), pp. 1202–1217.

[Fur06] P. G. Furth, B. Hemily, T. H. J. Muller, and J. G. Strathman: *TCRP Report 113: Using Archived AVL-APC Data to Improve Transit Performance and Management*. Washington, D.C., 2006.

[Gov11] R. M. P. Goverde and L. Meng: "Advanced monitoring and management information of railway operations". In: *Journal of Rail Transport Planning & Management* 1.2 (2011), pp. 69–79.

[Gov13] *Grenzeloze Openbaar Vervoer Informatie.* (Last Access: 18 July 2013). URL: http://govi.nu.

[Gvb13] *GVB maakt real-time reisinformatie open data*. (Last Access: 18 July 2013). URL: http://webwereld.nl/cloud/56325-gvb-maakt-real-time-reisinformatie-open-data.

[Hic04] M. HICKMAN: "Evaluating the Benefits of Bus Automatic Vehicle Location (AVL) Systems". In: D. LEVINSON and D. GILLEN (EDS.): *Assessing the Benefits and Costs of Intelligent Transportation Systems.* Boston: Kluwer, 2004, chapter 5.

[Kan08] E. M. KANACILO and N. VAN OORT: "Using a rail simulation library to assess impacts of transit network planning on operational quality". In: J. ALLEN, E. ARIAS, C. A. BREBBIA, C. J. GOODMAN, A. F. RUMSEY, G. SCIUTTO, and N. TOMII (EDS.): *Computers in railways XI*. Southampton, UK: WITpress, 2008, pp. 35–44.

[Kon02] A. KÖNIG and K. W. AXHAUSEN: "The Reliability of the Transportation System and its Influence on the Choice Behaviour". In: *Proceedings of the 2nd Swiss Transportation Research Conference*. Monte Verità, 2002.

[Oor09a] N. VAN OORT and R. VAN NES: "Control of public transport operations to improve reliability: theory and practice". In: *Transportation research record* 2112 (2009), pp. 70–76.

[Oor09b] N. VAN OORT and R. VAN NES: "Line length versus operational reliability: network design dilemma in urban public transportation". In: *Transportation research record* 2112 (2009), pp. 104–110.

[Oor10] N. VAN OORT and R. VAN NES: "The impact of rail terminal design on transit service reliability". In: *Transportation Research Record* 2146 (2010), pp. 109–118.

[Oor12] N. VAN OORT, J. W. BOTERMAN and R. VAN NES: "The impact of scheduling on service reliability: trip-time determination and holding points in long-headway services". In: *Public Transport* 4.1 (2012), pp. 39–56.

[Oor13] N. VAN OORT: "Incorporating enhanced service reliability of public transport in cost-benefit analyses". In: *Public Transport* 2013. – in press.

[Ope13] *openOV*. (Last Access: 18 July 2013). URL: `http://www.openov.nl/`.

[Rei13] *9292 gelooft in open data*. (Last Access: 18 July 2013). URL: `http://9292opendata.org/`.

[Spa13] D. SPARING and R. M. P. GOVERDE: "Identifying effective guaranteed connections in a multimodal public transport network". In: *Public Transport*. 2013. – in press.

[Str00] J. G. STRATHMAN, T. KIMPLE, K. DUEKER, R. GERHART, and S. CALLAS: "Service reliability impacts of computer-aided dispatching and automatic location technology: A Tri-Met case study". In: *Transportation Quarterly* 54.3 (2000), pp. 85–102.

[Waa88] J. VAN DER WAARD: "The relative importance of public transport trip time attributes in route choice". In: *Proceedings PTRC*. London, 1988.

[Vuc05] V. R. VUCHIC: *Urban Transit, Operations, Planning and Economics*. New Jersey: John Wiley and Sons, 2005.

[Wat03] B. WATERSON, B. RAJBHANDARY, and N. HOUNSELL: "Simulating the Impacts of Strong Bus Priority Measures". In: *Journal of Transportation Engineering* (Nov./Dec. 2003), pp. 642–647.

[Wil09] N. H. M. WILSON, J. ZHAO, and A. RAHBEE: "The Potential Impact of Automated Data Collection Systems on Urban Public Transport Planning". In: *Schedule-Based Modeling of Transportation Networks*. Springer, 2009, pp. 75–99. DOI: 10.1007/978-0-387-84812-9_5

[Xua11] Y. XUAN, J. ARGOTE, and C. F. DAGANZO: "Dynamic bus holding strategies for schedule reliability: Optimal linear control and performance analysis". In: *Transportation Research Part B: Methodological* 45.10 (2011), pp. 1831–1845.

Corresponding author: Niels van Oort, Delft University of Technology, Stevinweg 1, 2628 CN Delft, the Netherlands, phone: +31 6 1590 8644, e-mail: n.vanoort@tudelft.nl

Effects of Cooperative Traffic Signals on Tramway Operation

Christian Gassel, Jürgen Krimmling
Technische Universität Dresden

Abstract

Traffic lights have significant impact on traffic conditions at all, but also on energy consumption of approaching vehicles. The Dresden University of Technology is investigating several paths to reduce unnecessary stops at traffic lights. Finally, two energy saving strategies have been examined in depth: (I) Energy savings by traffic signal control and (II) by an onboard Driver Advisory System (DAS). As a result of research, a system was developed where cooperative traffic lights fulfil multi-model needs. Furthermore, the first DAS for energy efficient tramway control was designed, named COSEL. COSEL receives data about green phases from cooperative traffic lights and is already used in tramway operation.

Keywords: Energy efficiency, Tramway, Cooperative Traffic Lights, Traffic Management, ITCS

1 Introduction

Since automotive industry has been researching intensively on low-emission car mobility for some years, public transport has to compete with private transport on ecological aspects. In addition, energy prices tend to become more relevant for public transport operators.

At present industry and scientific institutions are researching intensively on energy storage systems for public transport, which increase the possibilities of recovering the kinetic energy of vehicles under regenerative braking. In contrast, this paper concentrates on operational measures to reduce tractive energy consumption already in advance.

In case of not having unlimited preemption, tractive energy consumption of tramways is influenced by traffic lights. Several strategies have been examined, which reduce unnecessary stops at traffic lights causing additional tractive energy consumption. This paper describes two approaches to reduce energy consumption of tramways in combination with traffic lights:

I – Energy savings by traffic light control

Energy consumption can be reduced by fitting the green phase to the arrival time of the approaching vehicle. The driver passes the traffic signal without being informed about green

phases. In this case, unlimited preemption for public transport represents the maximum reduction of unnecessary stops. However, unlimited preemption cannot be realized in each case due to conflicts of several vehicles approaching an intersection at the same time. The Dresden University of Technology developed a multi-criteria signal control (MCSC) system to decide about public transport preemption at traffic lights. That system significantly influences the stopping rate of public transport vehicles and is described in section 2.

II – Energy savings by DAS onboard

Driver advisory systems supporting energy efficient driving have already been developed for European railways and metro systems [Rai09]. These systems calculate optimal trajectories in real-time. The algorithms used differ in single-train and multi-train optimisation, single-section and multi-section optimisation as well as heuristic and numerical optimisation approaches [How95],[Liu03]. In railways DAS can provide in average energy savings of 5-10%. For tramways DAS for energy efficient driving never have been used in practise. The reasons are linked with the specific transport mode characteristics, e.g. vehicle actuated signal control with green phases, which are difficult to predict. In section 3 various onboard control measures for tramway operation are described to reduce unnecessary stops. Furthermore, the approach of COSEL (Computer Optimised Speed Control for Energy-efficient Light-rails) is presented. COSEL represents the first driver advisory system for tramways in regular operation.

2 Multi-criteria signal control

MCSC determines tramway preemption under consideration of operational issues (punctuality, headway regularity, connection services and optimal order of tramway vehicles at intersections, where the order can be influenced) and private transport issues (traffic flow). Therefore, MCSC evaluates all feasible green stages with configurable multi-criteria cost functions (see Formula 1).

Each green stage i gets costs on influencing deviation from schedule q_{FP} of public transport vehicles (Figure 1 left side). In addition, the influence on hindering motorized individual traffic is defined by q_{MIT}. (Figure 1 right side). The relationship of costs q_{FP} and deviation from schedule is modelled linearly. Green stages leading to high deviation are evaluated with higher costs, than green stages which reduce deviation. No costs are set for vehicles being slightly early or punctual.

The costs concerning individual traffic are composed of constant costs G_{MIT} and dynamic costs. Dynamic costs are calculated for each private traffic flow *j* which is influenced by the green phase. These costs q_{VL} depend on current LOS (Level of Service) of the specific traffic flow *j* and a weight *f* between public transport and private transport. The traffic flow with maximum costs is decisive.

Optionally, green stages also get costs at places where traffic lights may support connection services q_{Con} and where they are able to arrange an optimum order of vehicles q_{Or}. In general, q_{Con} and q_{Or} can only assume two different values. Either costs are set (goal is not fulfilled) or not (goal is fulfilled).

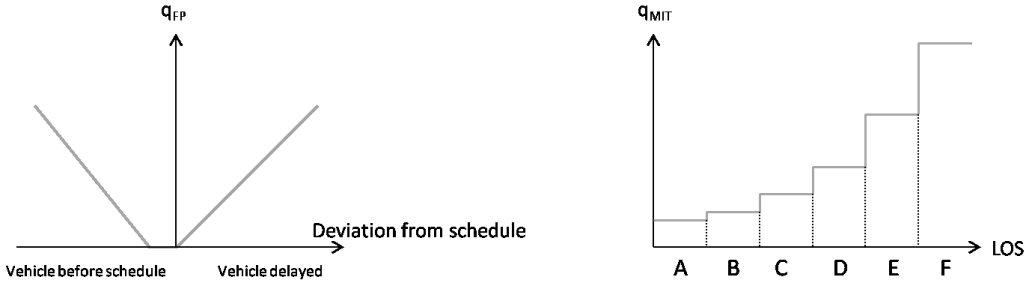


Figure 1: Consideration of private traffic flow (left) and deviation from schedule of public transport vehicles (right) in cost functions

Finally, MCSC chooses the optimum green phase Q^* with the minimum sum of multimodal costs.

$$Q^* = \min_{\forall i}\left(q_{\mathrm{MIT},i} + q_{\mathrm{FP},i} + q_{\mathrm{Con},i} + q_{\mathrm{Or},i}\right) \tag{1}$$

$$q_{\mathrm{MIT},i} = G_{\mathrm{MIT},i} + \max_{\forall j}\left(q_{\mathrm{VL},j} \cdot f_{ij}\right) \tag{2}$$

The stopping rate depends on the criteria mentioned above and the parameter settings used. However, the amount of energy savings ΔE by avoiding a stop at traffic lights is affected by several parameters (vehicles attributes, track infrastructure, energy supply management and operational restrictions). With respect to operational restrictions maximum speed limits v_{max} in the conflict area have intensive impact on energy savings in combination with stopping rates (Figure 2 right side).

In case of low speed limits (e.g. due to switches close to traffic lights) an additional stop has less impact on energy consumption, because of short-term acceleration activities to reach the speed limit again. Considering that the driver receives no information about green phases the vehicle is not approaching traffic lights with anticipating driving style.

Figure 2 illustrates on the left side the driving behaviour assumed with and without a stopping event. The dependency of energy savings and maximum speed restriction in the conflict area are shown on the right side. Calculations base on a vehicle mass of 41000kg. In this example gradients and recuperation rate while braking are assumed to be zero.

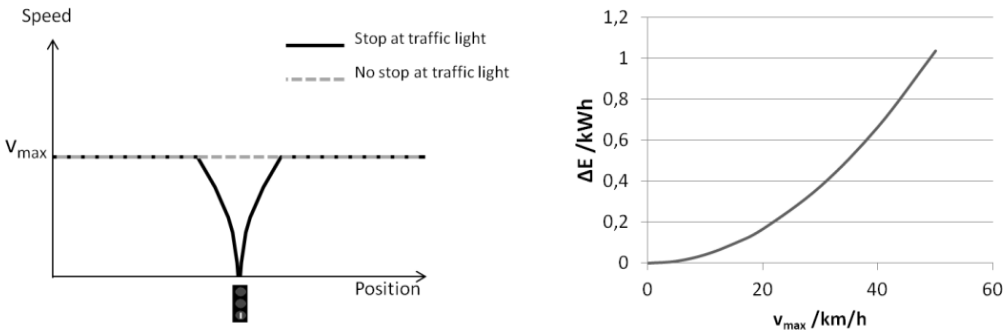


Figure 2: Driving trajectories (left) and energy savings avoiding one stop (right) using intelligent traffic light control

3 Onboard control measures

Various onboard control measures are feasible to reduce unnecessary stops in tramway operation. Basically, this paper concentrates on measures with cooperative systems (traffic lights or ITCS) which supply an onboard DAS with information about future conflicts areas. These conflict areas are signalled intersections as well as commercial stops being occupied by other vehicles where unnecessary stops may occur. However, the specific advice is derived on running time reserves based on optimal target times at conflict areas.

3.1 Optimum target time

In general, the optimum target time t_{opt} is an element of the set of target times T at the conflict (time span at the, f.i. green phases). Because transport capacity always plays a significant role at conflict points, the optimum target time firstly result from capacity consideration. In case of traffic lights, tramways usually should pass them as early as possible to reduce the influence of public transport on the whole intersection. At stopping areas, vehicles should arrive at stops as soon as the stop is not occupied anymore. In addition, these rules also fulfil the operational requirement of reduced journey times. However, due to safety reasons, the optimum target point at traffic lights is set some seconds later than the beginning of the green phase. Therefore, the driver approaching traffic lights has been aware of a green signal aspect for some seconds and is not concerned of approaching at red signals.

The optimum target time is shifted to a later point of time, not until the driver is unable to approach at the conflict area in time, even with time-optimal driving.

3.2 Dwell time control

Originally, the control of dwell times is an approach to guide the tramway driver to a target point without any influence the driving style between two commercial stops. The driver gets information about the optimum time of departure while waiting at the stop. For tramway movement toward the conflict area time-optimal driving is assumed. Figure 3 illustrates dwell time control on the left side. At the stop the advice “Drive” is shown as long as the driver would be able to reach the first green phase at the following traffic light (second 40 to 55 after the vehicle started at the previous stop). If the driver is not able to departure until second 20, the first green phase would be missed. Therefore, the system calculates another dwell time to reach the optimum target time $t_{opt,2}$ and shows a waiting advice with remaining dwell time.

That approach is already applied in practise of tramway operation, but not as solution on-board. Signals are positioned at the commercial stop to indicate the optimum departure time.

However, dwell time control being used solely is not able to influence a vehicle which has already departed. The energy efficiency only results from avoiding an unnecessary stop at the conflict area not from energy efficient driving towards the conflict area. In addition, conflicts have to be predicted at least when the vehicle stands at the previous commercial stop.

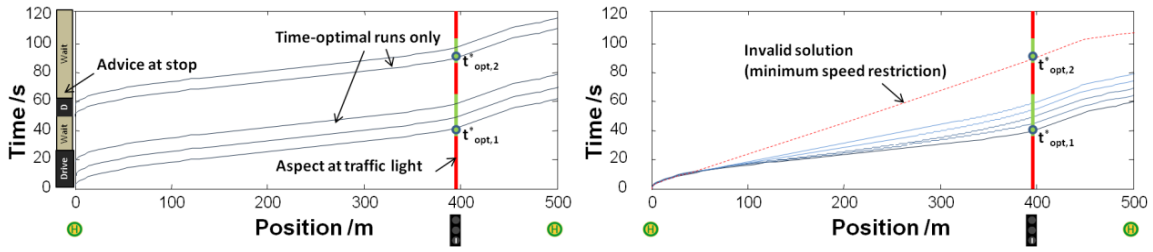


Figure 3: Difference between dwell time control (left) and speed control (right)

3.3 Speed control

In contrast, speed control is used to guide moving vehicles towards the optimum target point at the conflict area. In order to arrive with minimum energy consumption, the maximum principle of Pontryagin is used [Liu03]. It results in four regimes, which can occur in various sequences [Alb12]:

- acceleration with maximum permitted acceleration,
- cruising at constant speed,
- coasting without tractive energy consumption and
- braking with maximum permitted deceleration.

Multiple relevant parameters are considered in optimisation like vehicle attributes (e.g. driving dynamics) and track parameters (e.g. gradients, maximum speed profile).

In contrast to dwell time control, speed control not only saves energy by avoiding an additional stop. Additionally, energy is being saved by anticipatory driving when approaching a traffic light (for instance the coasting regime without tractive energy consumption). However, tramway infrastructure and operational issues result in limited applicability of speed control - especially in city centres where several public transport lines are being operated and public transport interferes with motorised individual transport waiting times at conflict areas strongly vary (0s-80s). Due to short distances between two consecutive stops (about 300m-600m) and operational constraints (minimum speed constraints and a very few changes of driving advice) these waiting times cannot be eliminated exclusively by energy efficient speed control. Figure 3 shows speed control on the right side. The optimum trajectory depends on the optimum target time at the traffic light. Though, the green phase being realised later (target point $t^*_{opt,2}$) cannot be reached because of minimum speed restrictions.

3.4 COSEL - Combination of dwell time control and speed control

The advisory system COSEL combines the advantages of both dwell time control and speed control. At the commercial stop the driver gets information about the optimum departure time. A countdown indicates the earliest departure time, which avoids approaching the

conflict area with cruising speed below minimum speed. Additionally, the driver gets information about the maximum dwell time which can be used for passenger interchange without taking the risk of missing the time window at the next conflict area. If the driver departures at the commercial stop COSEL will give speed advice.

Figure 4 illustrates several optimum trajectories depending on departure time at the previous stop. In the case shown, the driver has to wait 22 seconds at the commercial stop until a valid trajectory is given by DAS, which fulfils the minimum speed restriction and allows an arrival at the traffic light exactly at the optimal target time (dotted trajectory coloured light blue). If the driver waits longer because of passenger interchange, the running time reserves will decline, respectively approaching speed will increase. In case of not existing running time reserves, the optimal target time at the traffic light is shifted to a later point of time as long as the target time represents a part of the green phase (departure between second 45 to second 62). Thereby the driver is able to avoid a stop at the traffic light with time-optimal driving (trajectories coloured navy blue).

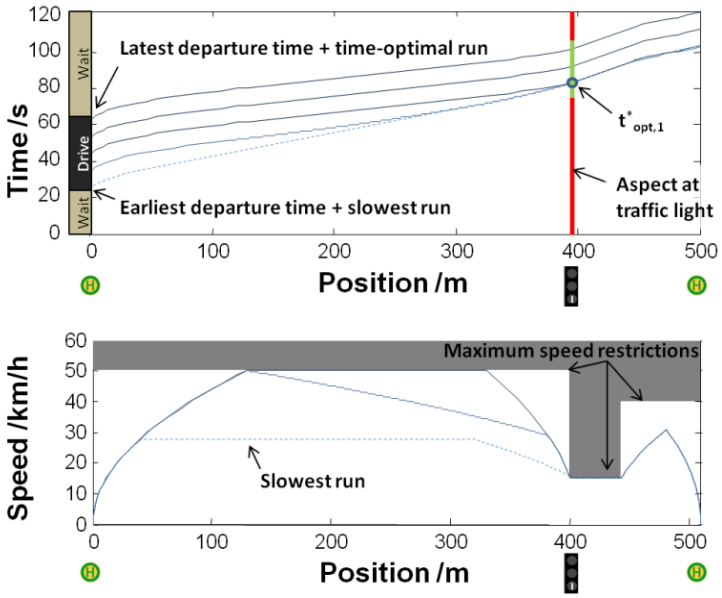


Figure 4: Optimum driving trajectories with respect to tramway specific restrictions

4 Implementation and experience of regular operation

Since July 2011 MCSC has been running in practical operation in Dresden, Germany. In order to use traffic data for multi-criteria optimisation MCSC receives current data from two central management systems. Firstly, the Dresden traffic management system VAMOS supplies information about road traffic conditions [Kre12]. Secondly, the Dresden ITCS being used by the Dresden Transport Operator provides specific operational data of public transport.

4.1 Data communication

In order to avoid an isolated solution data communication bases on several standardised interfaces being used in several European countries.

Approach I only requires unidirectional data transfer from the vehicle to MCSC. No data are sent to the vehicle. Since VDV (Association of German Transport Companies) defines various data interfaces/standards which are used in public transport systems all over Europe, MCSC is able to use a set of VDV interfaces. Both, operational issues of public transport considered in multi-criteria cost function and automatic vehicle location data, are derived by data already being contained in VDV 45x- as well as in VDV R09.16. Communication between MCSC and the local signal control unit bases on OCIT (Open Communication Interface for Road Traffic Control Systems). Using OCIT makes the approach usable for a wide range of signal control units (e.g. already tested with Siemens and Swarco).

Approach II additionally involves bi-directional data transfer from MCSC to public transport vehicles. Since data required are not covered by an existing standard interface the specific interface VLP100 (TU Dresden development) is used. Thereby, WLAN (IEEE 802.11n) is used as wireless networking standard during the pilot phase. With lessons learned from applicability of VLP100 further standardisation actions are planned.

4.2 Effects from traffic light control

MCSC influences traffic light control on a section with a total length of one kilometre, where two tramway lines (3, 8) and two bus lines (61, 66) are operating. On that section three intersections are controlled by MCSC (Nürnberger Platz, Fr.-Löffler-Straße/Reichenbachstrasse and Fr.-List-Platz). At these intersections unlimited preemption for both tramway lines cannot be applied in each case because of conflicts with the major road legs (for instance tramway routes at Nürnberger Platz have conflicts with average daily road traffic of 28,000 vehicles, in addition diversion route of the motorway nearby and coordinated green phases with intersections in neighbourhood). MCSC is evaluating road traffic conditions of the major road as well as operational aspects of the public transport.

Figure 5 demonstrates the influence of schedule delay on tramway preemption at the traffic signal Nürnberger Platz. As a result of cost function q_{FP}, MCSC tends to result in higher priority for trams being late, than trams which are on time or trams which are operating even too early. Thus, average time losses and the stopping rate are declining, when delay increases. Respectively, MCSC will be able to influence the energy consumption of a tramway system, even if the driver has no advisory system onboard (approach I). In comparison with common vehicle actuated control (CVAC) previously used at Nürnberger Platz, MCSC reduces the stopping rate from 75% to 46% (line 3, delay>2min). Because MCSC not only respects punctuality, further dependencies may occur, e.g. in case of good road traffic conditions (LOS A/B) tramway preemption can be higher – respectively the stopping rate even declines below 40%.

4.3 Effects from driver advisory system

In addition, the Dresden Transport Operator (Dresdner Verkehrsbetriebe AG) has been applying the driver advisory system COSEL since June 2012 (approach II).

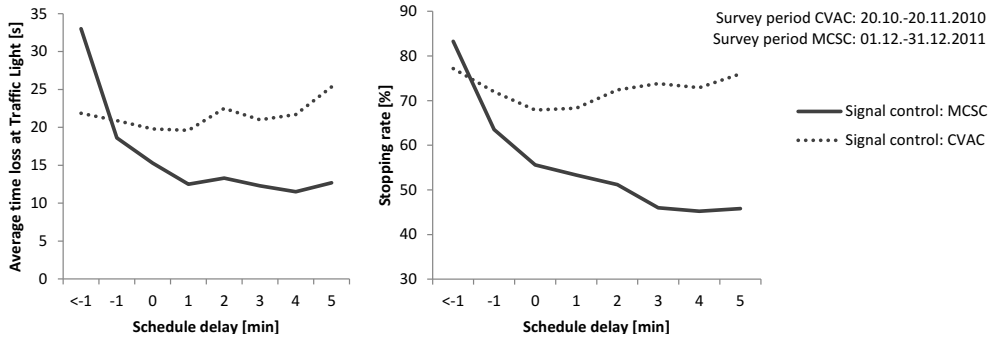


Figure 5: Average time loss (left) and stopping rate (right) depending on deviation from schedule at Nürnberger Platz (Line 3 direction Coschütz, approach I without DAS)

For the first time such a DAS has been designed for tramways, supporting drivers to avoid unnecessary stops at traffic lights. Therefore MCSC receives status data of cooperative traffic lights as well as operational data from the ITCS (Intermodal Transport Control System). In addition, green phases at traffic lights are chosen by MCSC (similar to approach I) and driving trajectories of other vehicles are predicted to estimate occupation times at stops. These spatial and temporal data about expected conflicts are sent to the DAS onboard. Finally, COSEL calculates the optimum driving trajectory ensuring smooth driving without unnecessary operational stops.

Since, such an advisory system never has been used in regular tramway operation before, several migration stages have been arranged to prove the effects on energy consumption. In summer 2012 twelve voluntary drivers evaluated the system performance of COSEL in regular operation. The results of several hundred runs per month document the drivability of given advice by COSEL in urban tramway environment.

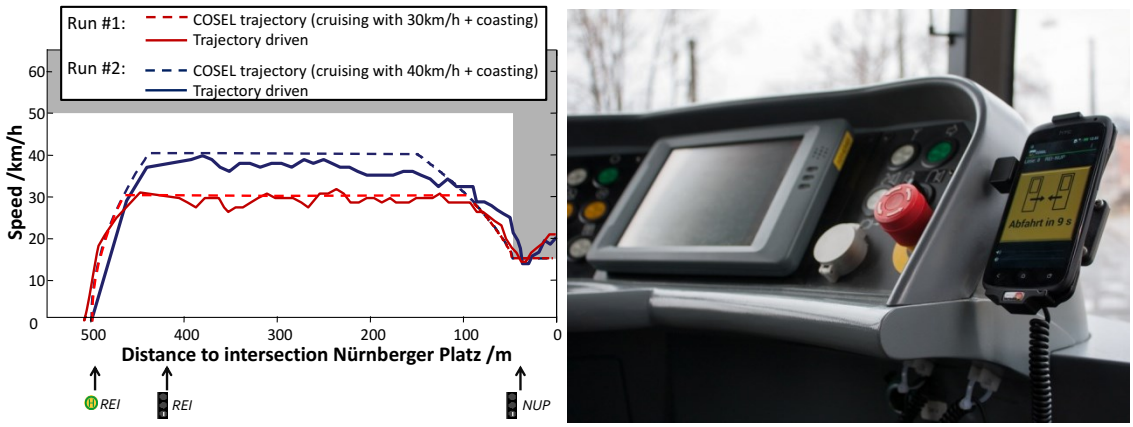


Figure 6: Driver advisory system COSEL (left: two trajectories recommended and actual driven in regular operation; right: integration of COSELmobile in the driver's cabin)

Figure 6 (left side) illustrates two runs between stop Reichenbachstrasse and stop Nürnberger Platz. COSEL calculates optimum trajectories depending on green phases of two traffic lights running independently of one another. Even if gradients strongly vary on this section (0 to 4%) the drivers are able to follow advice and finally avoid stops at traffic lights. For the sake of high acceptance the experiences of the drivers have also been implemented in

software updates. As long as COSEL is not implemented on an OBU (On-Board-Unit) the software runs on a smartphone (COSELmobile). The device is installed in the driver's cabin (see Figure 6 right side).

The average stopping rate of tramway vehicles being equipped with COSEL declined to 24% (Table 1). Despite the fact that COSEL only gives advice, the system actually has an impact on the driver's behaviour, respectively the stopping rate of the vehicles. Some tramway vehicles still stopped because of technical reasons, which cannot be detected automatically (e.g. defect infrastructure components, being relevant for operation) and human reasons, like deviating driving styles from recommendation to test possibilities and limitations of that new system.

Table 1: Effects on average stopping rate of line 3 direction Coschütz at Nürnberger Platz

Case	Survey period	Average stopping rate
CVAC	10.2010-11.2010	0.70
MCSC (Approach I)	12.2011	0.52
COSEL (Approach I+II)	06.2012-08.2013	0.24

The effects on energy consumption depend on the distribution of running time reserves, acceptance rates, recuperation rates and system availability. For data analysis, power data recorded by the Dresden Measuring Tram are used [Har09]. According to that, tractive energy consumption was reduced by 4-12% per run on the section analysed (between stop Reichenbachstrasse and stop Nürnberger Platz). In summary, operational measures like intelligent traffic control and driver advisory systems are able to make additional potentials accessible to reduce energy consumption in tramway networks. These measures have to be considered as a part of a bundle of measures on various fields of research, which complete each other (e.g. energy recovery systems and operational measure described in this paper) to maximise energy efficiency of public transport at all.

Having tested the system successfully with voluntary drivers, the Dresden Transport Operator arranged the next migration stage in autumn 2013, involving all tramway drivers. Therefore, the system was installed on 83 tramway vehicles (about 50% of the whole tramway vehicle fleet), which are operating in more than 90% of all cases on the sections being supplied with traffic light information for COSEL.

5 Future actions

In 2013/14 approach II will be applied on one of the heaviest charged transport corridors in Dresden. By then, MCSC will control traffic lights on at total length of 4.2 kilometres, where the headway of tramways partly falls below two minutes during rush hour. Moreover, COSEL will be implemented as a module of OBU. In this regard, data transmission will not base on WLAN standard anymore, but on an existing digital data transmission network owned by Dresdner Verkehrsbetriebe. In addition, a third approach will implemented, which also considers energy savings in MCSC. The cost function of MCSC is extended by energy/fuel consumption of public and private transport vehicles. The green phase with minimum costs

will be calculated by MCSC and transmitted to the DAS onboard. Algorithms used by DAS equal to approach II.

Acknowledgement

The authors would like to thank the Urban Road Department of Dresden (Straßen- und Tiefbauamt Dresden) and the Dresden Transport Operator (Dresdner Verkehrsbetriebe AG) for hosting this research and valuable support while development and implementation.

References

[Alb12] T. Albrecht, A. Binder, and C. Gassel: "Applications of real-time speed control in rail-bound public transportation systems". In: *IET Intelligent Transportation Systems*. 2012. doi: 10.1049/iet-its.2011.0187.

[Har09] M. Harter, M. Beitelschmidt, G. Striegler, and I. Sauermann: "Die Dresdner Messstraßenbahn – Konzept, Architektur, Komponenten". In: *ETR – Eisenbahntechnische Rundschau* (Dec. 2009).

[How95] P. Howlett and P. Pudney: *Energy-efficient train control.* Berlin: Springer, 1995.

[Kre12] A. Kretschmer and J. Krimmling: "The traffic management system VAMOS – from research to regular operation". In: *19th World Congress and Exhibition on Intelligent Transport Systems and Services*. Vienna, Austria, 2012.

[Kri10] J. Krimmling and C. Gassel: "Interfacing ITCS and road traffic control system – a synthesis to increase quality and LRT energy efficiency". In: *10th UITP Light Rail Conference*. Madrid, Spain, 2010.

[Liu03] R. Liu and I. M. Golovitcher: "Energy-efficient operation of rail vehicles". In: *Transportation Research Part A* 37 (2003).

[Mat11] T. Matschek, C. Gassel, and J. Krimmling: "Cooperative traffic lights under consideration of the needs of public transport and motorized individual transport". In: *8th European Congress on ITS.* Lyon, France, 2011.

[Rai09] Rail Safety and Standards Board Ltd.: *Driver advisory information for energy management and regulation.* Stage 1 Report, 2009.

Corresponding author: Christian Gassel, Technische Universität Dresden, "Friedrich List" Faculty of Transport and Traffic Sciences, Institute of Traffic Telematics, 01069 Dresden, Germany, phone: +49 351 463 36749, e-mail: Christian.Gassel@tu-dresden.de

Pre-signals for Bus Priority: Basic Guidelines for Implementation

S. Ilgin Guler, Monica Menendez
ETH Zurich

Abstract

Buses operating mixed with cars can often get stuck in car congestion. One commonly used solution is to dedicate a lane for bus use only. However, when bus flows are low, dedicated lanes running through intersections can reduce the discharge flows from these locations and lead to increased car delays. Therefore, a commonly used solution is to discontinue the dedicated lane upstream of the main signal. In this paper, we advocate the use of pre-signals at these locations to continue providing bus priority while minimizing disruptions to car traffic. Pre-signals can allow buses to jump the car queues upstream of signalized intersections, while allowing cars to utilize the full capacity of the main signal when buses are not present.

The goal of this paper is to provide a basic summary of previous research done on pre-signals. Using this information, ideas on how to operate pre-signals and practical guidelines on how to implement them at signalized intersections will be provided. Lastly, insights on when pre-signals can reduce the system wide (buses and cars) person hours of delay will also be given. This information can then be used to determine where and when pre-signals should be implemented in real urban networks.

Keywords: Pre-signals, Bus priority, Traffic flow, Signalized intersections

1 Introduction

Reliable and fast public transportation (i.e., buses) is an important tool to relieve urban areas of car congestion. However, the operation of buses in urban environments can often be impeded by interactions with cars. Bus operations can become slow and unreliable when buses are caught in car queues. Also, car operations can be hindered by buses which stop frequently and pull in and out of bus stops. As a result, both car and bus modes operate less efficiently and the capacity for both can be reduced.

A commonly used solution is to dedicate a lane for bus use only. Bus lanes have been studied as early as 1975 [Lev75]. However, theoretical analysis of roadway capacities with

the use of bus lanes has only been recently done. These lanes can allow the buses to bypass car queues in order to reduce travel time and increase their reliability. Nevertheless, their effect on the system-wide (i.e., buses and cars) person hours travelled can differ greatly especially based on the bus frequency.

When bus frequencies are high, dedicated bus lanes can also benefit cars by reducing the number of conflicting maneuvers experienced between the two modes. If these dedicated lanes are run through bottleneck locations the discharge flows can even be expected to increase [Men07]. However, if bus frequencies are very low and the bus lanes are severely underutilized, then even the increased discharge flows from the adjacent lanes might not be enough to compensate for the wasted space. In this case, bus lanes will lead to additional car delays due to reduced capacities at intersections, and also longer queue lengths since there is one less lane available for cars to queue on. The longer queues can also increase the risk of queue spill overs to other intersections and further reduce discharge flows.

In order to mitigate the negative effects of bus lanes on the system, dynamic use of bus lanes have been proposed on links and at intersections. On signalized arterials, intermittent bus lanes were proposed [Vie01, Vie04]. Intermittent bus lanes allow cars to use all lanes on a roadway except for when buses are present. Whenever buses are arriving, a space in front of the bus is dedicated for bus use only allowing the bus to always pass without interfering with cars. However, since buses receive priority on the roadways only when they are present, cars can fully utilize the capacity of the roadway when there are no buses. A similar idea, named bus lanes with intermittent priority, was proposed by Eichler and Daganzo, 2006 [Eic06]. Guler and Cassidy, 2012 [Gul12] theoretically determined the bounds for which dynamically shared lanes would increase the total capacity of the system. Intermittent bus lanes were field tested in Lisbon, Portugal [Vie07], and in Melbourne, Australia [Cur08]. These experiments found a varying range for bus delay reductions.

With a special focus on intersections, the use of pre-signals, which are additional signals upstream of the main signal, was proposed by Wu and Hounsell, 1998 [Wu98]. The goal was to increase discharge flows from intersections while still providing bus priority. That paper provides three different operating strategies for the pre-signal and theoretically evaluates the delays associated with each. A similar idea, but with a slightly different operating strategy, was proposed by Guler and Menendez, 2013 [Gul13a]. They analyzed the car and bus delays encountered at pre-signals when the main signal is under saturated [Guler13a], and when the main signal is over saturated [Guler13b].

There are only a few places where pre-signals have been implemented in the real world. These locations include London, U.K. [TFL05], and at least one implementation in Zurich, Switzerland. In London, Transport for London provides guidelines on how to assess and design pre-signals. This document qualitatively talks about the different points to consider when assessing the feasibility, design and implementation of pre-signals. However except for a general suggestion for the distance from the intersection of a pre-signal location, quantitative assessment tools are not provided in this document [TFL05]. The car and bus delays encountered at the Zurich pre-signal were empirically evaluated [Gul13c], and used to validate the theoretical analysis presented in [Guler13a].

The goal of this paper is to provide a comprehensive view on the operation of pre-signals. This paper presents basic guidelines on how and when pre-signals can be implemented in urban environments. These guidelines are drawn from the analytical findings of Guler and Menendez [Guler13a, Guler13b, Guler13c]. However, the results have are presented to provide practitioners with tools to assess the feasibility of pre-signals and present basic guidelines for design. To do so, Section 2 will describe the operation of pre-signals. Section 3 will discuss the conditions for which pre-signals can improve the system for under saturated and over saturated intersections. Finally, Section 4 will offer some final remarks for the implementation of pre-signals.

2 Operation of pre-signals

A pre-signal is used when there are 2 or more lanes approaching the intersection in the same direction as the bus, and one is dedicated for bus-use only. The idea is to discontinue this bus lane some distance upstream of the intersection, and use a pre-signal to stop cars at this location in order to still provide bus priority at the main signal (see Figure 1). The pre-signal allows cars to use all lanes to discharge from the main signal, except when a bus arrives. At that time, the pre-signal turns red for cars. This allows buses to maneuver into the intersection without encountering conflicts from cars, and provides bus priority by allowing them to jump the car queue present upstream of the pre signal (Guler and Menendez, 2013a).

Pre-signals operate as described below [Guler13a]. This described operation provides the smallest combined bus and car delays. This statement is justified later.

- The pre-signal turns red for cars:
    - In advance of a red main signal such that cars queue only upstream of the pre-signal. This ensures that an arriving bus can move to the stop line at the main signal and discharge immediately when the main signal turns green.
    - When a bus arrives to the pre-signal, regardless of the main signal's phase. This gives the bus priority to move to the main signal without encountering conflicting maneuvers from cars.
- The pre-signal turns green for cars:
    - In advance of the green main signal such that no gaps are created in the car discharge from the main signal. This implies that if the number of car lanes upstream of the pre-signal is less than the number of mixed use lanes (as is in Figure 1), cars discharging from the pre-signal would briefly form a queue again at the main signal to not starve the main signal of flow.
    - When a passing bus clears the pre-signal (given that pre-signal would have been green if no bus were present).

Now imagine that the duration of red at the pre-signal is different than described above. If the red time at the pre-signal is shorter, this would imply that a longer queue would form at the main signal every cycle. This is true since now there exists a longer duration between the green at the pre-signal and at the main signal. In this case, a bus arriving during the red main signal, but green pre-signal would queue behind more cars, leading to a larger average bus delay. The cars would experience no benefit from this type of operation since they would still be able to discharge from the main signal at the same time.

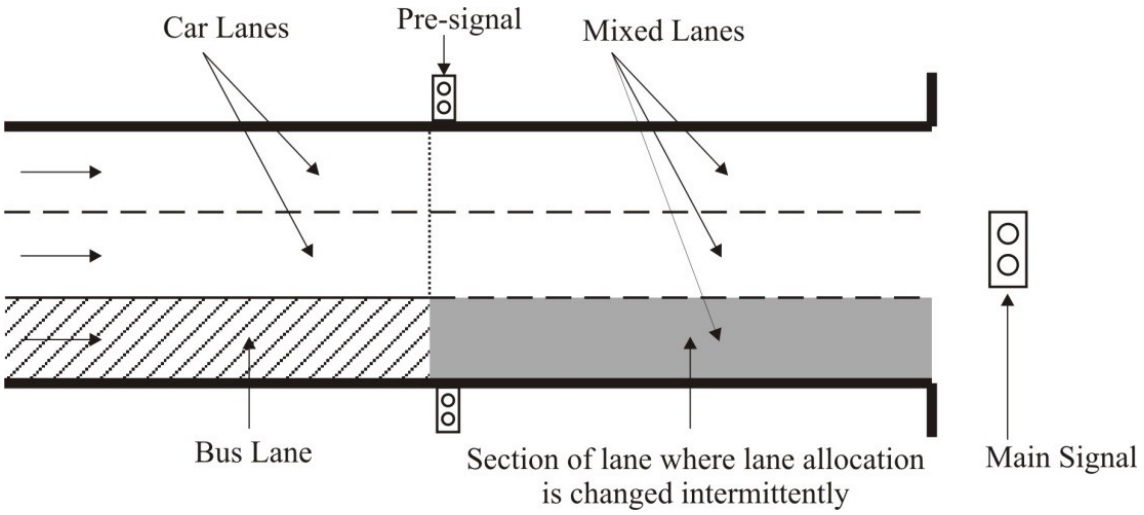


Figure 1: Schematic of a pre-signal [12].

Now imagine the opposite, where the red at the pre-signal is longer than described above. In this case, cars could not discharge from the main signal at saturation rate for as long as they could if the above described operation was used. This is true since the pre-signal has less number of lanes than the main signal. In this case, cars would experience additional delays. While the buses could benefit from this operation since the queue of cars at the main signal would be shorter, further investigation of this strategy showed that this reduction in delay was not enough to compensate for the increase in car delay.

The queuing of cars at the pre-signal and at the main signal for the operation described above when buses are not present can be visualized in Figure 2. This figure represents a signalized intersection where the cycle length is C hours, the red duration at the main signal is r_{ms} hours, and the red duration at the pre-signal is r_{ps} hours. The virtual arrival of cars to the main signal is shown as line A with slope a veh/hour. The virtual arrival of cars to the pre-signal is shown as line D_{ps} with a maximum slope of s' veh/hour. Here s' represents the capacity of all the car lanes upstream of the pre-signal. The departure of cars from the main signal is shown as line D_{ms} with a maximum slope of s veh/hour. Here s represents the capacity of all lanes at the main signal. Notice that $s \geq s'$ since there are more (or the same number of) lanes present at the main signal as compared to the pre-signal.

As can be seen in Figure 2, cars initially start queuing at the pre-signal during the pre-signal red time. Then the pre-signal turns green in advance of the main signal (i.e., $r_{ps}<r_{ms}$) and cars are required to queue at the main signal for a short duration. This happens since typically there is one less lane available at the pre-signal and the discharge flow is not

enough to saturate the green at the main signal. Thus, cars queue at the main signal in order to fully utilize its capacity. However, this does not imply that cars experience additional delays. For a cycle during which buses are not present, the car delay is exactly the same as it would be if the pre-signal did not exist, even though the number of stops is increased for some cars. For a cycle during which a bus is present, cars will experience some additional delays.

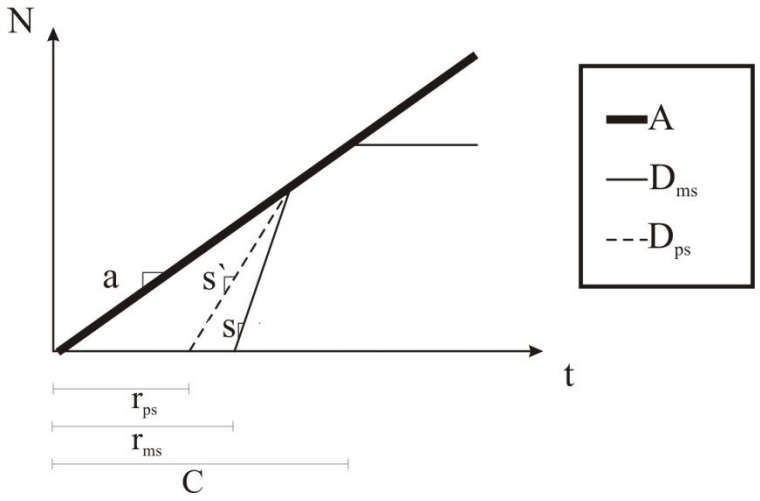


Figure 2: Queuing diagram for an under saturated signalized intersection with pre-signals when buses are not present [Gul13a].

For a perfectly saturated main signal (when the demand is equal to the capacity of the signal) there needs to be enough space between the pre-signal and the main signal such that the main signal can be fully utilized by the queued cars (when *s>s'*). Therefore, the section of the lane where the lane allocation is intermittently changed, needs to extend at least for a distance of *d* km (see equation below). In other words, the pre-signal must be located at least at a distance *d* upstream of the stop bar at the main intersection.

$$d = (C - r_{ms}) \cdot s \cdot \frac{1}{k_{jam}} \tag{1}$$

where k_{jam} is the jam density of all lanes at the main signal in veh/km. In the case when *s=s'*, *d* is only limited by the space required for the bus to be able to maneuver into the appropriate lane.

The next step is to determine the duration of red at the pre-signal (r_{ps}). Note that Figure 2 shows the virtual departure curve from the pre-signal. This implies that the departures from the pre-signal have been shifted by the free flow travel time between the main signal and the pre-signal (d/v_f, where v_f is the free flow speed). Hence, in real implementations the pre-signal turns red in advance of a red main signal by the free flow travel time. Then the duration of r_{ps} can be found as:

$$r_{ps} = \frac{s \cdot (s' - a)}{s' \cdot (s - a)} \cdot r_{ms} \quad (2)$$

Since r_{ps} depends on the demand rate, it could either be pre-determined by using historical data or determined dynamically by measuring the arrival rate with the use of loop detectors.

In the case when s=s', the pre-signal would still turn green d/v_f in advance of a green main signal, but r_{ps} would be equal to rms. Hence, the first car queued at the pre-signal would arrive to the main signal just as it was turning green and would not be required to stop again. Therefore, no secondary queue of cars would form at the main signal. The pre-signal would turn red again d/v_f in advance of a red main signal

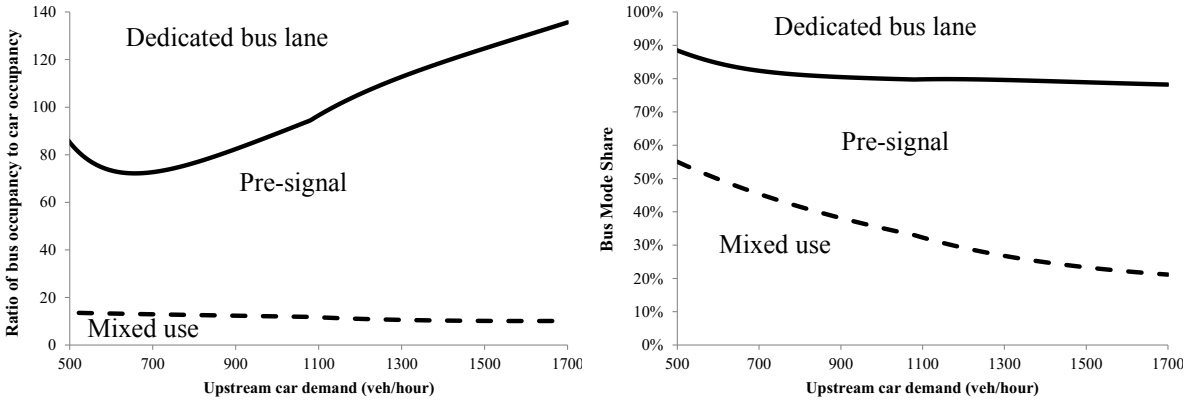
If a bus arrives when the cars have a green pre-signal, the pre-signal needs to turn red for cars for a certain duration r_b. This duration needs to be enough to allow the bus to maneuver into the appropriate lane downstream of the pre-signal. This duration is typically around 5 seconds, however, it could be as high as 10 seconds as it was observed at the Zurich pre-signal.

3 Bounds of application for pre-signals

This section presents the bounds for which pre-signals can provide the lowest system-wide (car and bus) total person hours of delay. These bounds represent when the system will benefit more by the use of pre-signals compared to mixed use lanes (where only cars would benefit) and dedicated bus lanes (where only buses would benefit). These bounds, for which pre-signals outperform the two other strategies, are calculated using two metrics: (*i*) ratio of bus occupancy to car occupancy; and, (*ii*) mode share of buses assuming a car occupancy of 1 for the intersection of interest (notice that this is the most conservative scenario possible).

Figure 3a shows the ratio of the bus occupancy to the car occupancy, and Figure 3b shows the mode share of buses for which each strategy would provide the lowest system-wide delay for a range of car demands at under saturated intersections. Above the solid line, dedicated bus lanes perform the best; between the solid line and the dashed line, pre-signals perform the best; and below the dashed line, mixed use lanes perform the best. These results are based on the analytical formulations presented in Guler and Menendez [Gul13a]. However, the analytical formulations of [Gul13a] have been extended to also look at mode share as a decision metric. As can be seen in Figure 3a, pre-signals can improve the system for a wide range of bus occupancies. As the upstream car demand increases these bounds become wider. Overall, pre-signals always outperform the two other strategies when bus occupancies are between 18 and 70 times greater than car occupancies. Figure 3b shows the same bounds using the bus mode share of only the specific intersection considered as the metric of analysis. Similar to Figure 3a, the domains of application of pre-signals increase (i.e., the difference between the solid and dashed lines become wider) as the car demand increases. For pre-signals to be the best for the system, buses are required to carry between 60-90 % of the passengers when car demands are low, but only 20% (still up to 90%) of the passengers

when car demands are high. Notice that a high bus mode share for the specific intersection is not representative of the mode share of the entire city. The mode shares in Figure 3b are representative of intersections with high bus flows, where a high number of people would be expected to be traveling by bus. It is unclear how the entire city's mode share would be reflected in these calculations. However, it is expected that the bus mode share for the city would be much lower than at the specific intersection considered.



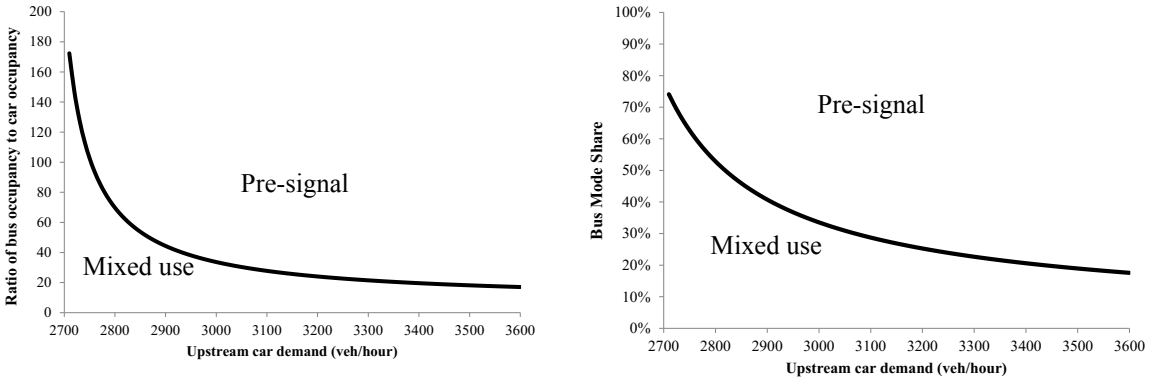
(a) (b)

Figure 3: Bounds on (a) occupancy [Gul13a], and (b) bus mode share for which pre signals reduce total person hours of delay at under saturated intersections [s = 5400 veh/hr, s' = 3600 veh/hr, r_{ms} = 40 sec, C = 80 sec, additional red time at pre-signal = 5 sec, headway = 80 seconds].

Figure 4 shows a similar analysis for over saturated signals [Gul13b] assuming that there exists a non-peak hour demand equal to 85% of the main signal's capacity. Figure 4a shows the ratio of the bus occupancy to the car occupancy, and Figure 4b shows the mode share of buses for which each strategy would provide the lowest system-wide delay for a range of car demands at over saturated intersections.

Notice that, with such a high non-peak demand, the queues would never clear and delays would be infinite if dedicated lanes were used. Even if the non-peak demand was 0, for dedicated bus lanes to improve the system the bus occupancies would need to be at least 1000 times greater than car occupancies, and bus mode shares would need to be ~100%. This result directly follows from the discussion in the introduction. When the signal becomes over saturated, dedicating a lane for bus use only significantly reduces the discharge flow from the main signal and the car delays increase very rapidly. Therefore, even if there are many passengers in the bus, this does not compensate for the additional delay all cars incur.

Based on that, only the comparison between pre-signals and mixed use lanes is non-trivial in the case of an over saturated intersection. As can be seen in Figures 4a and b, when the intersection is barely over saturated mixed use lanes are better for the system for realistic bus occupancies. However as the car demands increase further the pre-signals start becoming more competitive with mixed use lanes.



(a) (b)

Figure 4: Bounds on (a) occupancy [Gul13a], and (b) bus mode share for which pre signals reduce total person hours of delay at over saturated intersections [s = 5400 veh/hr, s' = 3600 veh/hr, r_{ms} = 40 sec, C = 80 sec, additional red time at pre-signal = 5 sec, headway = 80 sec].

Combining the results of Figures 3a and 4a, Figure 5 can be obtained as a summary of the range for which pre-signals benefit the system as compared to mixed lanes. The bus occupancies for which pre-signals provide the lowest system delays are quite low except when the main signal is slightly over saturated. For a small range of demands (between 100% and 105% of the main signal's capacity) the bus occupancy needs to be >100 times greater than a single car's occupancy.

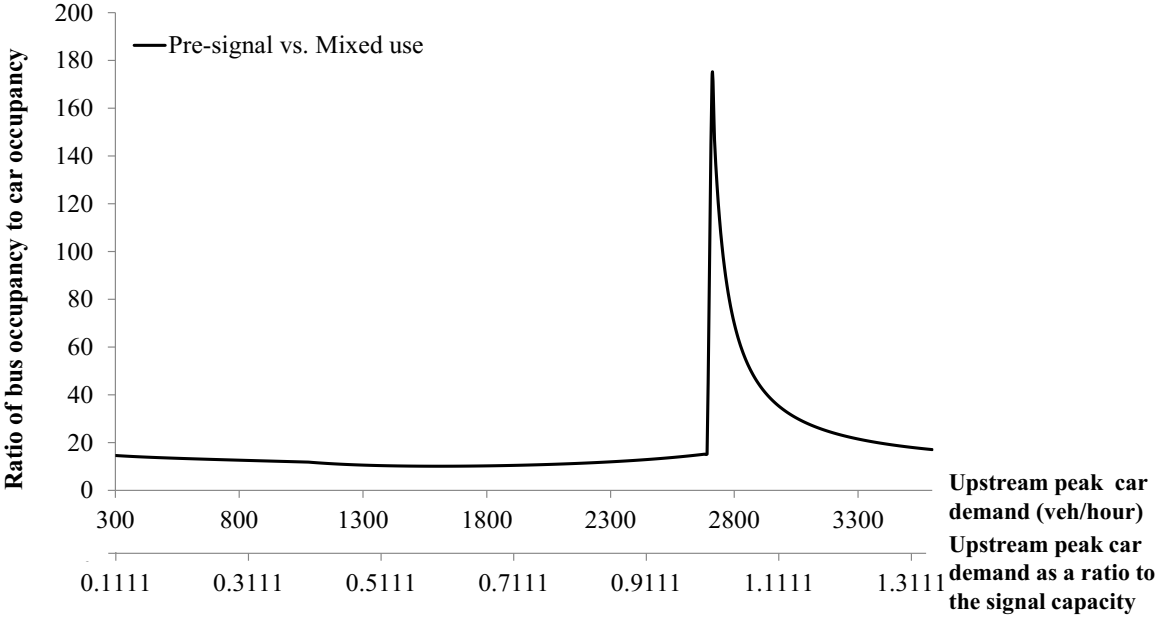


Figure 5: Bounds on occupancy for which pre-signals reduce total person hours of delay for under saturated and over saturated intersections [s = 5400 veh/hr, s' = 3600 veh/hr, r_{ms} = 40 sec, C = 80 sec, r_b = 5 sec, h = 80 sec] [Gul13b].

In light of these findings, it is important to take precautions in the operation of pre-signals to always minimize the system-wide delays. To improve the traffic operations and avoid causing excessive system-wide delays the pre-signal could be turned off (i.e., turned green for both buses and cars) when the traffic operations at the main signal are detected to approach over saturated conditions. The criteria for turning off/on a pre-signal could be determined in advance by considering demand as a fraction of the main signal's capacity (e.g., as demand approaches 95% of the signal capacity the pre-signal could be turned off). This would allow buses to always jump the car queues present, at least until they reach the location of the pre-signal. Notice, however, that the additional benefits gained by buses from the presence of the pre-signal would not be observed while the pre-signal is turned off. In other words, turning off the pre-signal would cause additional delays to buses, but when demand is close to the main signal's capacity, it can also keep the main signal from becoming over saturated and avoid excessive car delays. The pre-signal could be turned back on whenever the demand decreases or increases further. Again, the criteria for how large the demand needs to be can be determined in advance as a fraction of the signal capacity (e.g., as demand reaches 110% of the signal capacity the pre-signal could be turned on again). According to Guler and Menendez, [Gul13b], pre-signals can benefit the system for reasonable bus occupancy values if the car demand is smaller than the main signal capacity, or 1.05 times greater than the main signal capacity. The examples described above used 95% and 110% to include some buffer for demand variations.

Note that the bounds above are shown for headways of 80 seconds. These are relatively short headways (high bus frequencies) which would only be observed along busy bus corridors where perhaps multiple lines meet. For longer headways, the bounds would become lower making pre-signals even more attractive compared to the two other bus-car operating strategies. Hence, the bounds shown in the above figures can be considered as the most conservative scenario possible. To determine bounds for a specific location with specific input values, the equations in Guler and Menendez [Gul13a, Gul13b] can be used.

4 Conclusions

This paper provided a summary of the analytical evaluations of pre-signals. The results show that for realistic bus occupancies and mode shares, pre-signals can provide the lowest system-wide delays. Especially at key intersections within an urban environment where car delays are large, rather than running bus only lanes through the signalized intersection, pre-signals can be seen as a good compromise that benefits both buses and cars. A word of caution is necessary here since, if the main signal is barely over saturated, pre-signals could increase the system-wide delays. Therefore, when implementing these strategies in real life, the car demand should be closely monitored, so when demand gets close to the main signal's capacity the pre-signal can be turned off. This will allow for buses to still have priority over cars, at least until the location of the pre-signal, but can also maximize the discharge flow from the main signal avoiding additional delays.

Even though there are cases where pre-signals are not the best for the system, in most scenarios they do perform better than dedicated bus lanes. Since they result in similar bus

delays, and also improve bus reliability at many signalized intersections, the use of pre-signals should be considered as an alternative to dedicated bus lanes running through intersections.

In comparison with mixed use lanes, pre-signals evidently provide higher bus reliability, and lower delays, potentially encouraging mode changes. In addition, they are more conducive to implementing transit signal priority, especially in over saturated conditions. The arrival time of the bus to the main signal can be predicted with greater accuracy, and the bus is at the head of the main signal in more situations as compared to mixed lanes. Thus, more accurate changes can be made to the main signal timings to further reduce the bus delays.

Looking at the analytical evidence presented in this paper, pre-signals can be beneficial to the system over a range of situations. The number of implementations of these strategies can be widely extended to provide bus priority in a politically feasible fashion. By improving bus service, this mode can be made more attractive to users in order to induce mode changes. These mode changes could further reduce car delays and queues. Overall, this then could lead to more efficient and sustainable transportation systems in urban environments.

References

[Eic06] M. EICHLER and C. F. DAGANZO: "Bus lanes with intermittent priority: Strategy formulae and an evaluation". In: *Transportation Research Part B* 40.9 (2006), pp. 731–744.

[Gul12] S. I. GULER and M. J. CASSIDY: "Strategies for sharing bottleneck capacity among buses and cars". In: *Transportation Research Part B* 46.10 (2012), pp. 1334–1345.

[Cur08] G. CURRIE and H. LAI: "Intermittent and dynamic transit lanes: Melbourne, Australia, experience". In: *Transportation Research Record: Journal of the Transportation Research Board* 2072 (2008), pp. 49–56.

[Gul13a] S. I. GULER and M. MENENDEZ: "Analytical formulation and empirical evaluation of pre-signals". In: *Transportation Research Part B* (2013) – under review.

[Gul13b] S. I. GULER and M. MENENDEZ: "Evaluation of pre-signals at over saturated signalized intersections". In: *93rd Annual Meeting of the Transportation Research Board.* Washington, D.C., 2013 – submitted.

[Gul13c] S. I. GULER and M. MENENDEZ: "Empirical evaluation of bus and car delays at pre-signals". In: *The 13th Swiss Transport Research Conference (STRC)*. 2013.

[Lev75] H. S. LEVINSON, C. ADAMS, and W. F. HOEY: *NCHRP Report 155: Bus Use of Highways Planning and Design Guidelines*. TRB, National Research Council, Washington, D.C., 1975.

[Men07] M. MENENDEZ and C. F. DAGANZO: "Effects of HOV lanes on freeway bottlenecks". In: *Transportation Research Part B* 41.8 (2007), pp. 809-822.

[TFL05] TRANSPORT FOR LONDON: *Bus pre-signal assessment and design guide*. (Last Access:

Jan 2009). URL: www.tfl.gov.uk/assets/downloads/businessandpartners/Bus_Pre-Signals_Aug05.pdf.

[Vie01] J. VIEGAS and B. LU: "Widening the scope for bus priority with intermittent bus lanes". In: *Transportation Planning and Technology* 24.2 (2001), pp. 87–110.

[Vie04] J. VIEGAS and B. LU: "The intermittent bus lane signals setting within an area". In: *Transportation Research Part C* 12.6 (2004), pp. 453–469.

[Vie07] J. VIEGAS, R. ROQUE, B. LU and J. VIEIRA: "The intermittent bus lane system: demonstration in Lisbon". In: *86th Annual Meeting of the Transportation Research Board*. Washington, D.C., 2007.

[Wu98] J. WU and N. HOUNSELL: "Bus priority using pre-signals". In: *Transportation Research Part A* 32.8 (1998), pp. 563–583.

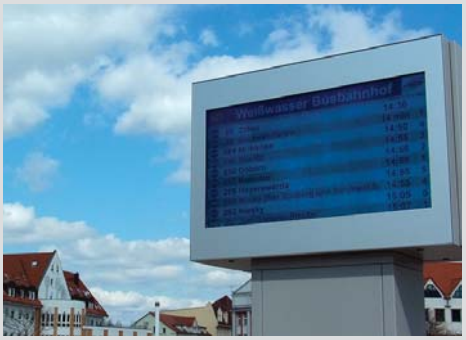
Corresponding author: S. I. Guler, ETH Zurich, Institute for Transport Planning, 8093 Zurich, Switzerland, phone: +41 44 633 3192, e-mail: ilgin.guler@ivt.baug.ethz.ch

Know-how und Leidenschaft für nachhaltige Mobilität

VCDB VerkehrsConsult Dresden-Berlin GmbH bietet Beratung, Planung und Service zu allen Fragen und Problemen des Verkehrswesens.

Dabei sind die in der VCDB GmbH umfangreich vorhandenen Betriebserfahrungen die Grundlage für praxisgerechte und zukunftsweisende Ergebnisse in den Bereichen

- **Fahrzeugtechnik**
- **Verkehrsplanung**
- **Verkehrstechnik**
- **Verkehrstelematik**
- **Elektromobilität**
- **Betriebsassistenz**



Das Team Verkehrstelematik der VCDB berät Verkehrsunternehmen, Verkehrsverbünde und Hersteller bei Planung, Einführung und Betrieb von Telematiksystemen im öffentlichen Personenverkehr.
Unsere Schwerpunkte dabei sind:

- **Rechnergesteuerte Betriebsleitsysteme**
- **Digitale Kommunikationssysteme**
- **Systeme zur Dynamischen Fahrgastinformation**
- **Zugleit- und Sicherungssysteme**

Wir erstellen Analysen, Systemkonzepte, Wirtschaftlichkeitsuntersuchungen sowie Spezifikationen (Lastenhefte) und übernehmen das komplette Projektmanagement von der Ausschreibung bis zur Systemabnahme. Darüber hinaus bieten wir kundenspezifische Hard- und Softwarelösungen für Lichtsignalbeeinflussung und Fahrgastinformation an.

Weitere Informationen? Besuchen Sie uns auf **www.vcdb.de**!

VCDB VerkehrsConsult Dresden-Berlin GmbH

Traffic Signal Preemption by Means of Digital Transmission Methods

Michael Preusker[1], Charlotte Gäbel[2], Stefan Löwe[1]

[1] VerkehrsConsult Dresden-Berlin GmbH (VCDB)

[2] Technische Universtät Dresden

Abstract

The transmission of messages from Public Transport (PT) vehicles to traffic light system still takes place in an analogue manner. In this paper, alternative digital transmission methods will be presented. These methods are analysed in order to determine whether they meet the requirements of PT operation in terms of a practice-capable release time correctness and reliability of the transmission.

Here, the focus is on the use of the data service (GPRS / UMTS) in the public mobile radio systems. Since regional PT vehicles are often equipped with an Automatic Vehicle Location (AVL) on-board computer and a wireless GPRS / UMTS modem, it is examined how suitable this existing equipment is for realizing the quality oriented influencing of traffic signals.

By appropriate measures, such as changing the position of the reporting point and time-delay compensation of transmission by timestamp method, the effects of inbuilt latencies should be compensated. In this paper the results of a pilot project to influence traffic signals via GPRS / UMTS in Dresden's regional PT are presented and evaluated.

Keywords: Traffic Signal Preemption, Digital Transmission, Prioritisation of Public Transport, Acceleration of Public Transport

1 Introduction

The acceleration of Public Transport (PT) vehicles at traffic light systems in urban areas with dense traffic on the one hand increases the attractiveness of PT, for instance by shorter travelling times and punctual departure times at stops on the other hand it improves the efficiency of operations, for instance a lesser vehicle demand to provide the same offer,. For acceleration measures, the communication between the PT vehicle and the light signal system is indispensable – in order to obtain an optimal phase by a timely pre-assignment , as well as to keep the obstruction of other traffic participants as low as possible by timely clear assignment. In Germany, the major part of message transmission from PT vehicles to a traffic

light system is still carried out by means of analogue radio frequency (RF) data transmission based on standards of the early 1980s. Although digital message transmissions are used increasingly, for instance on the basis of TETRA and Tetrapol, they constitute, due to a lack of standardization, spot solutions which are not compatible with each other. At present, public mobile radio networks are used in initial pilot projects. The aim of this paper is to show the digital alternatives to the conventional transmission of messages in PT by means of analogue radio systems, with special focus on the specific features of digital data transmission.

2 Basics of digital Transmission

2.1 Overview of mobile radio

For the transmission of voice and data from and to mobile terminals (MT), mobile radio networks are required, which provide access to services and applications independently of time and place. These MT are connected to the telecommunication infrastructure (TCI) by radio via an air interface. In their radio cell, base stations (BS) receive and transmit the signals for the corresponding terminal and forward them to the wire-bound TCI. The TCI consists of a hierarchically designed structure of various components. This structure serves for the control and organisation of the connection as well as for a guaranteed exchange of information. Mobile radio networks could be designed for a wide range of customers with a public access network, on the one hand, for instance in the form of GPRS and UMTS networks. On the other hand, there could be a corporate need for a separate closed network. For this application, the professional mobile radio (PMR) is used, on the basis of the TETRA standard, for instance. [Sau11]

2.2 Operational modes in digital radio

Operational modes indicate the connection between the MT and thus determine the use of frequency resources. They are divided into

- Trunked Mode (TM)
- Direct Mode (DM)

In this case, TM represents the standard operation of digital radio. Since the MT are communicating via a network infrastructure, this operational mode is used for both mobile radio networks of public digital radio and PMR. The range depends on the area coverage of the BS.

Alternatively, the DM provides a direct connection between the MT, without the integration of any network infrastructure. Basically, this allows for shorter latency times. Unlike TM, DM has only a local range, which depends on the topography and the electromagnetic nature of the environment. Repeaters can be used to increase the range of reception. [Ber03]

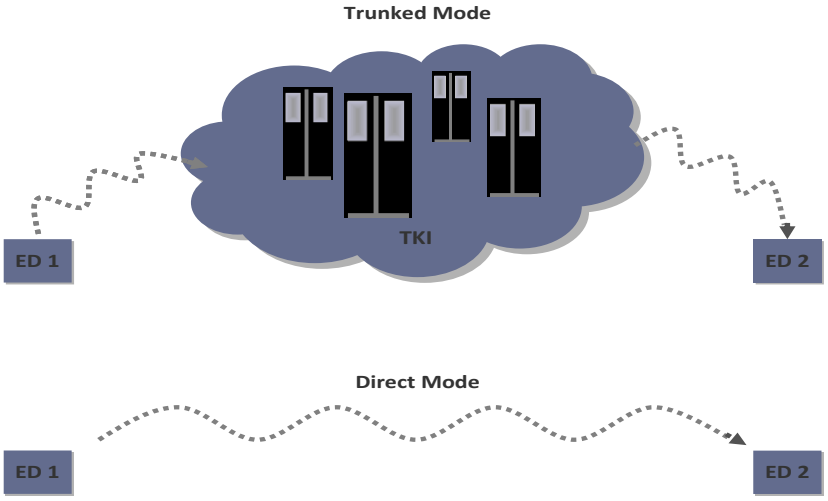


Figure 1: Operating Model.

2.3 System examples

Professional mobile radio (PMR)

PMR is a closed system (defined user group), which is used by public authorities and emergency service for voice and data communication, for instance.

Compared with public digital radio, the group of users is smaller, and the communication behaviour is characterised by shorter conversation times and a smaller number of connections. The signal delay of the connection is shorter. Examples for PMR systems are TETRA, Tetrapol and DMR.

Public mobile radio

The public mobile radio systems like GSM/GPRS and UMTS were developed for an information exchange of a wide customer range. The MT communication runs via a public accessible TCI, which the network operator places at the disposal of the user. It is an open system which provides a network infrastructure for the transmission of voice and data. It is more accessible than PMR, with less technical expenditure.

With the agreement “Third Generation Partnership Project” (3rd GPP), compatibility between the systems GPRS, UMTS and LTE has been brought about. Since the user behaviour is highly volatile, latency time is longer and less continuous than in PMR. [Sau11]

Local networks

For the established reporting point method Wireless Local Area Networks (WLAN) could be used as an alternative detection method. Concerning the detection method WLAN could determine the position and the moving direction of the vehicle. By means of signal transfer delay and signal strength, software could evaluate the distance to an access point. For more detailed information on this topic please refer to further reading.

Sensor networks

The use of sensor networks on the basis of the IEEE 802.15.4 standard is another new approach for the positioning of PT as well as special vehicles nearby traffic lights. They constitute an independent local network. The vehicle is equipped with a mobile sensor (reflector). The mast of a traffic light is equipped with a sensor (coordinator) which is connected to the traffic light system controller via an interface. This sensor introduces the phase-based distance measurement to the mobile reflector, which allows to ascertain the vehicle position.

3 Basics V2I in public transport (PT)

The PT vehicles communicate with the infrastructure to a very large extent. This concerns both the own non-public infrastructure, such as the point controls of rail vehicles and the public infrastructure, such as traffic light systems.

As a rule, a three-step method with the following message types is used for traffic light system influence:

- Pre-assignment (about 300 to 500 m before traffic light) [1]
- Main-assignment (about 100 to 150 m before traffic light) [1]
- Clear-assignment (about 5 m after traffic light) [1]

[1] Distance details are determined individually for each traffic light system and travel relation within the framework of a traffic engineering examination and could in individual cases deviate from the values mentioned.

It is also possible to use a four-step method with an additional transmission of the readiness of departure (doors closed, Stop-assignment). In simple cases, it is also possible to reduce the number of steps to two (Main-assignment, Clearassignment).

The message telegrams transmit at least the following information from the vehicle to the traffic light system:

- reporting point (contains the requested traffic flow, message type)
- timetable situation

Additional information can be transmitted [Lem13]:

- priority (between PT vehicles)
- line number
- course number (vehicle identification)
- destination and
- train length.

4 Application examples in public transport (PT) operation

4.1 Terrestrial trunked radio (TETRA)

TETRA is an open standard of PMR. For civilian use in PT, TETRA can be used for the transmission of messages to the traffic light controller.

For the transmission, TETRA is using the Time Division Multiple Access (TDMA) method. Therefore, data from various MT can be transmitted in one channel. In this process, several users share the same transmission frequency. By means of timeslots, which recur in fixed time intervals, the users are separated. Each MT has its own timeslot for the transmission of the individual data packets, which run through the shared channel consecutively. [Bül09], [Wal00]

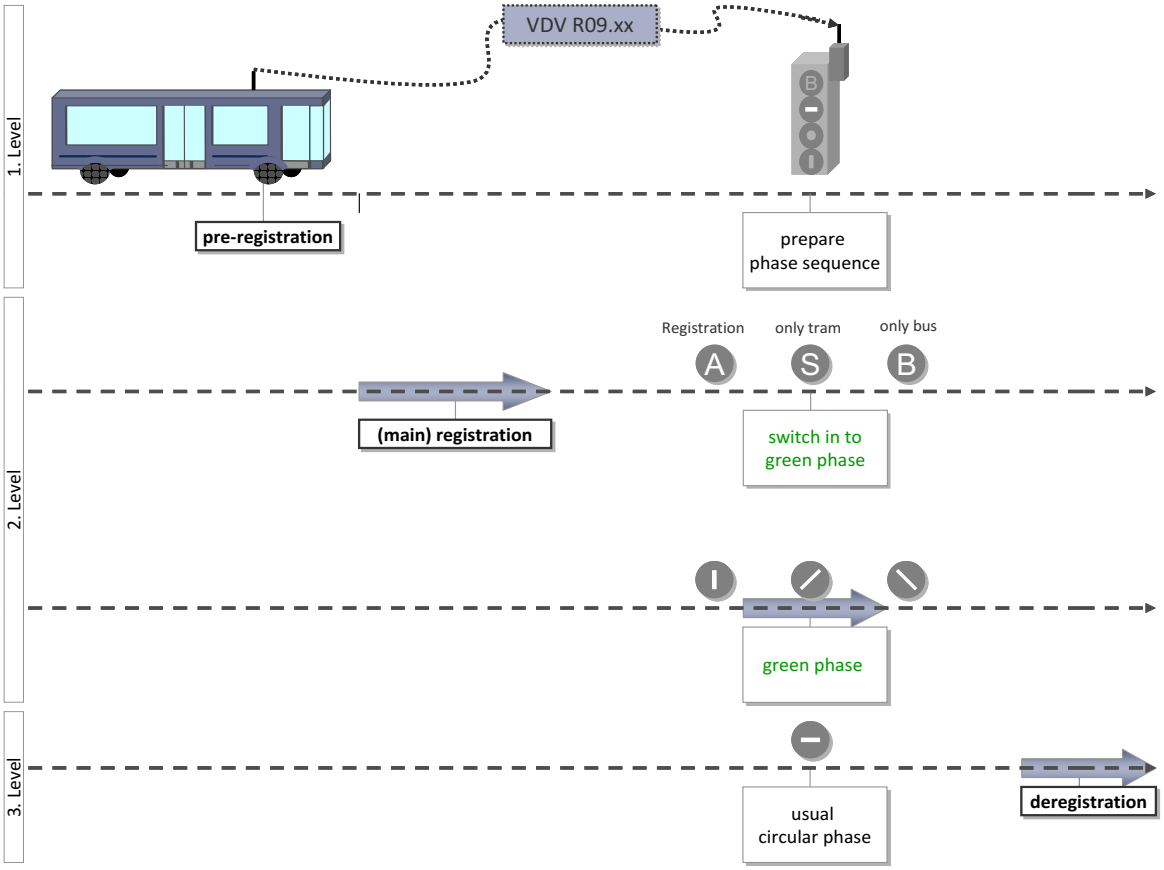


Figure 2: Message Point.

Because of the allocation of the timeslots, synchronisation must be performed in case of a change-over between DM and TM, which restricts the availability in voice communication too heavily. For this reason, TM is given preference in TETRA for traffic signal preemption.

In [Lem13] are listed two examples. The transport Köln PT company (KVB) uses the TETRA digital radio for the implementation of traffic signal preemption. Because of longer latencies compared to the analogue radio, the positions of the reporting points are adapted. The TETRA digital radio is also used by the east Ruhr PT cooperation (KöR). Their method is

based on the analogly minted traffic signal preemption too. In both cases the telegram is transmitted in the trunked mode (TM).

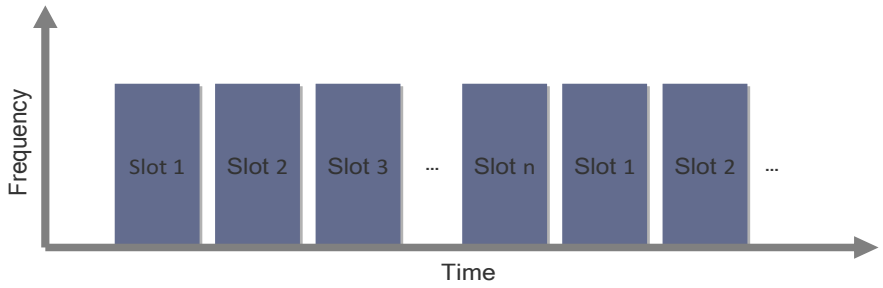


Figure 3: TDMA.

4.2 Tetrapol

Tetrapol is a non-open standardised digital and cellular trunked radio system which uses the multiple access method Frequency Division Multiple Access (FDMA). In this system, several users share an available frequency band. This frequency band is decomposed into many single narrow frequency bands and allocated to the transmitting MT. [Bül09], [Wal00]

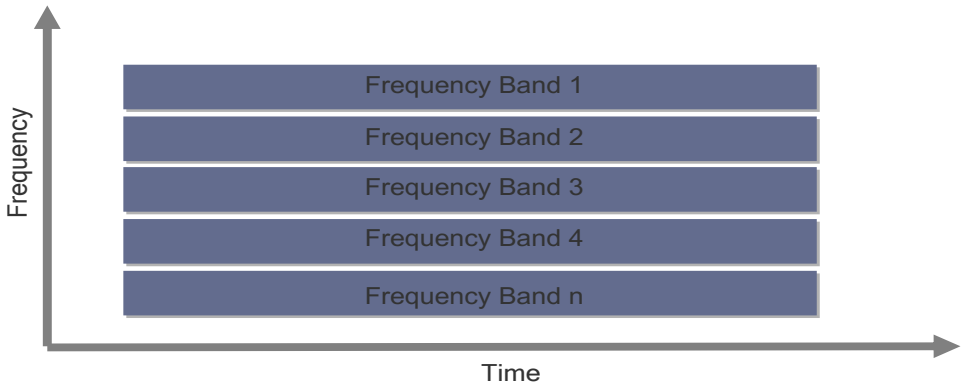


Figure 4: FDMA.

As this method does not require any time-consuming synchronisation like the TDMA method in TETRA, the change-over between TM and DM can be performed quickly. Therefore, it can use the advantage of a transmission without TCI and thus release faster traffic signal preemption.

The Berlin PT company (BVG) uses Tetrapol in the direct mode (DM) with an exclusive simplex channel. Considering the reporting point method telegrams are sent directly to the traffic light system. [Lem13]

4.3 GPRS / UMTS

An alternative for traffic signal preemption is the use of a public mobile radio network. Via a radio module fitted to the vehicle, which is implemented with the Automatic Vehicle Location

(AVL) on-board computer, for instance, the telegram can be transmitted via the public digital radio system to the PT receiver of the traffic light controller which is also to be equipped with a corresponding module to this end. This constitutes integration into the existing reporting point method for traffic signal preemption. Transmission in the public mobile radio system is always designed to take place via a TCI. A communication under DM is not envisaged. Since the data packets run through a TCI of complex design, and since they must be routed within the network, there are latencies which depend on the rate of utilisation of the BS, among other things. In addition, it should be noted that public mobile radio network operators proceed to a prioritisation of the data streams, which is normally oriented towards end users (Internet, Streaming, …) and which handle sporadic short data packets like in V2I with a lower priority

As a result, the latency is correspondingly discontinuous and is situated between 250 ms and 2,000 ms in GSM/GPRS, for instance. This discontinuity must be compensated by using an appropriate method.

4.4 Sensor networks

The phase-based distance measurement for positioning offers an alternative to the established reporting point method. It could be used independently of the available on-board computer.

Because of the continuous distance measurement, the previous fixed reporting point based method can be replaced by a system offering a considerably higher precision. Thus, the vehicle could be supervised during the entire run of the train, and fault situations (traffic congestion) can be recognized more easily.

However, it is not possible to transmit all those contents (lines, deviation, …) which are transmitted when using the reporting point method. As a result, another type of traffic engineering planning is to be taken into consideration.

5 Evaluation of pilot projects with different systems for traffic signal preemption

5.1 Latency as a comparison parameter

The Quality of Service (QoS) fixed in the International Telecommunications Union (ITU) -Z Recommendation E.800 defines the properties of a qualitative and efficient network, characterizing the transmission speed, the latency time and the susceptibility, for instance. In order not to distort the release time at the traffic light signal, the focus is on latency.

In the DM, the latencies are so small that they can be neglected.

On the contrary, as a result of the data transfer via a TCI, there are latencies in the TM which must be taken into consideration. The extent can be analyzed with the use of a time stamp in the telegrams. This examination results in the development of methods which are able to compensate a longer duration of transmission. [Sau11], [Sch12]

5.2 GPRS/UMTS

At present, initial field tests are conducted under the pilot project for “Regio-RBL Oberelbe” By means of statistical analyses, the records of the telegrams are studied in order to check the latency and the reliability of transmission with regard to the requirements of a timely traffic light system influence. In this connection, a time stamp method is taken into consideration, for instance, where the vehicles and the infrastructure must be equipped with synchronous clocks.

In order to compensate the discontinuous latencies, a telegram must be transmitted earlier and delayed on the receiver side according to the actual latency, before it is handed over to the traffic light system controller. To this end, the reporting point must be advanced in dependence of the maximum running time to be expected and of the vehicle speed.

Detailed results will be available at the MT-ITS.

5.3 Sensor networks

The pilot project is also used to check the necessary requirements on latency and accuracy for the application in PT.

Initial results will be available at the MT-ITS.

6 Summary and outlook

In the course of digitalisation, the previously dominating analogue transmission methods used in PT will also be replaced gradually. The traffic signal preemption is an important radio system for operation, because fast handling processes ensure reduced waiting times at intersections or, when leaving loading islands, provide for shorter running and turnover times. Transmission methods based on public digital radio allow for communication with the traffic light system, independently of the set-up of an own TCI. Especially for regional traffic, this is a cost-effective possibility of communication with the traffic light system. A low-expenditure technology is based on sensor networks the potential of which could be developed to bring about a method for traffic signal preemption.

References

[Ber03] F. BERGMANN, H. GERHARDT, and W. FROHBERG: *Taschenbuch der Telekommunikation*. Fachbuchverlag Leipzig, 2003.

[Bül09] F. BÜLLINGEN and P. STAMM: *Mobilfunknetze für professionelle Anwendungen*. Wik-Consult GmbH, 2009.

[Lem13] C. LEMENT, D. KLEIN, F. J. SENF, and B. RADERMACHER: *VDV Schrift 426, Lichtsignalanlagen-Beeinflussung über digitale Funktechnik*. Verband Deutscher Verkehrsunternehmen VDV, 2013.

[Sau11] M. SAUTER: *Grundkurs Mobiler Kommunikationssysteme*. Vieweg, 2011.

[Sch12] A. SCHILL: *Verteilte Systeme*. Springer Vieweg, 2012.

[Wal00] B. WALKE: *Eignung der Standards ETSI/TETRA und TETRAPOL zur Erfüllung der betrieblich-taktischen Forderungen der Behörden und Organisationen mit Sicherheitsaufgaben (BOS)*. Lehrstuhl für Kommunikationsnetze Aachen, 2000.

Corresponding author: Michael Preusker; VCDB VerkehrsConsult Dresden-Berlin GmbH; 01067 Dresden, Germany, phone +49 351 482 3117, e-mail: m.preusker@vcdb.de

Interfacing Conflict Resolution and Driver Advisory Systems in Railway Operations

Birgit Jaekel, Thomas Albrecht
Technische Universität Dresden

Abstract

For the integration of Conflict Resolution Systems and Driver Advisory Systems the paper presents the concept of train path envelopes. The Conflict Resolution System determines the order of trains for each section along a line and their routes through the network. Based on this information feasible time and speed intervals are computed and form the envelope, which can be used for optimisation of train driving without impact on capacity. The paper describes different optimisation stages: first the energy optimal paths for all trains are computed, then the buffer times are allocated. A case study is presented for freight train optimisation on a single-track line.

Keywords: Conflict Resolution, Driver Advisory System, Energy Consumption, Capacity, Optimisation, Pertubation Management, Railway

1 Introduction

1.1 Motivation

The European research project ON-TIME (Optimal Networks for Train Integration Management in Europe) aims at improving capacity in railway networks by introducing technologies for real-time traffic control [OTP13]. Conflict Resolution Systems (CRS) and Driver Advisory Systems (DAS) are seen as key technologies in this field.

A summary of the development of CRS has been given e.g. in [DAr08]. The most recent approaches perform train re-routing and replanning times of train services close to real time. Resulting outputs are in most cases train sequences on critical infrastructure elements (or on all elements), planned stops and routes. These systems are supposed to be integrated in track-side Traffic Control Centres in the next few years.

DAS were originally developed to support the driver in energy-efficient operation, but the use of DAS also enables the train driver to reach planned running times more exactly. Different systems are operational and described in the literature, see [Mit09] for an overview.

Recently, developments have been made to integrate CRS and DAS in order to e.g. avoid passing restricted signal aspects or stopping in front of red signals and thereby further increase capacity, energy efficiency and driving comfort. A few systems are already operational on networks of reduced complexity [Lag11; Mon09]. There is some literature on the subject of how additional target points (of times and velocities for specific positions) or time windows – speed and time restrictions for passing a certain location – can be used in the process of trajectory optimisation, e.g. [Alb12; Pud11].

This paper focusses on how target windows for DAS can be computed based on the result of CRS, which to the knowledge of the authors has not been described in the literature before. The sequence of target windows along a train run shall be called train path envelope. Train path envelopes shall be computed in such a way, that they enable maximal throughput through the bottlenecks, and – whenever possible – ensure maximum freedom for energy-efficient driving.

1.2 Drivability of Train Path Envelopes

The physical ability of a train to drive within a given envelope is called drivability of an envelope, it can be determined using microscopic train simulation.

The drivability of an envelope mainly depends on the given times and the velocity restrictions on the track. For a maximal flow at bottlenecks the velocity of the trains has to be close to the design speed of the infrastructure [Alb12], the ability to reach it depends mainly on train traction and track characteristics.

In Fig. 1 a sequence of time windows along a train run – a train path envelope – is illustrated. In order to reach the time window between sections s_2 and s_3 the train needs to drive quite slowly (close to the slowest possible running time without stop on this section which is illustrated with the dashed line). The next time window (between sections s_3 and s_4) cannot be reached even with time-minimal driving on section s_3 as can be seen from the dash-dotted line which arrives just after the end of the time window, therefore the train path envelope as a whole is not drivable.

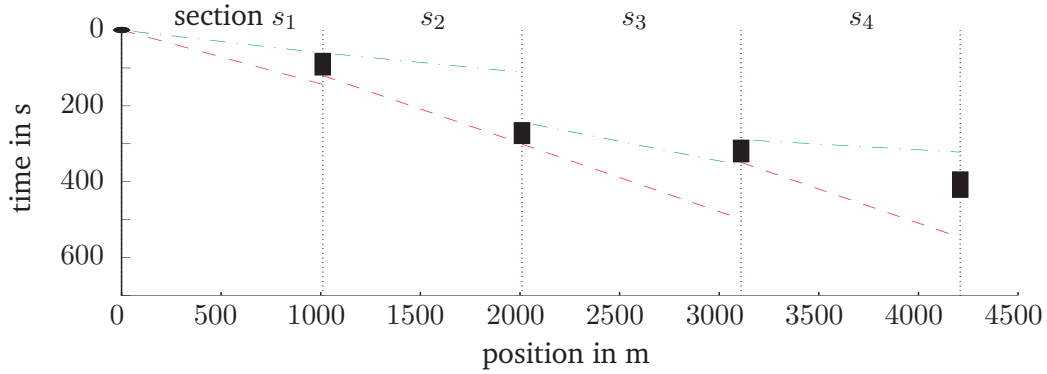


Figure 1: Non drivable envelope; boxes: envelope time windows, dashed: slowest path without stop on each section, dash-dotted: fastest path on each section.

2 Computing Envelopes with Potential for DAS Energy Optimisation

It should be noted, that some CRS use static train running times for train movement prediction due to the NP-hardness of the problem. As shown in [Sob11], this hardly influences the quality of the solution of CRS in terms of overall delay, but the times given as output of CRS (passing times, blocking times) might actually not be drivable. Therefore it was decided to use only a part of the CRS output for train path envelope computation, that is the train routes, train sequence over track sections and the information on planned stops. Train path envelope computation then takes place in two stages:

1. a drivable path is computed for each train, so that energy consumption for all trains becomes minimal.

2. the available times between each two consecutive train paths are then allocated to the train paths in order to obtain the envelope.

The remainder of this section explains in detail first the system model used and then each of the two optimisation steps.

2.1 System Model

As train service times on track sections are to be optimised there is a need to define the variables

$$\text{trains } \theta_i \in \Theta (i = 1 \dots n), \tag{1}$$

$$\text{sections } s_j \in S (j = 1 \dots m). \tag{2}$$

The occupation of a section s_j by a train θ_i is called event e_{ij}. The following relations are defined for section occupations of trains:

$$\text{relation } R_E = \{(i,j) \mid \theta_i \text{ operates on } s_j\} \tag{3}$$

$$\text{relation } <_{\theta,i} = \{j \leq_i k \mid e_{ik} \text{ is planned to take place before } e_{ij}\} \tag{4}$$

$$\text{relation } <_{s,j} = \{i \leq_j l \mid e_{ij} \text{ is planned to take place before } e_{lj}\} \tag{5}$$

$$\text{relation } R_{\theta,i} = \{(j,k) \in S^2 \mid j \leq_i k \wedge \nexists\, s_p \in S : j \leq_i p \wedge p \leq_i k\} \tag{6}$$

$$\text{relation } R_{s,j} = \{(i,l) \in \Theta^2 \mid i \leq_j l \wedge \nexists\, \theta_p \in \Theta : i \leq_j p \wedge p \leq_j l\} \tag{7}$$

Relation R_E holds all planned events. $R_{\theta,i}$ is the relation of all events of a single train θ_i which follow directly one after the other and $R_{s,j}$ is the relation of the events of a single section s_j which follow directly one after the other. The sets $(\Theta, S, R_E, <_{\theta,i}, <_{s,j}, R_{\theta,i}, R_{s,j})$ will be given by CRS. Fig. 2 gives an example of these sets and relations.

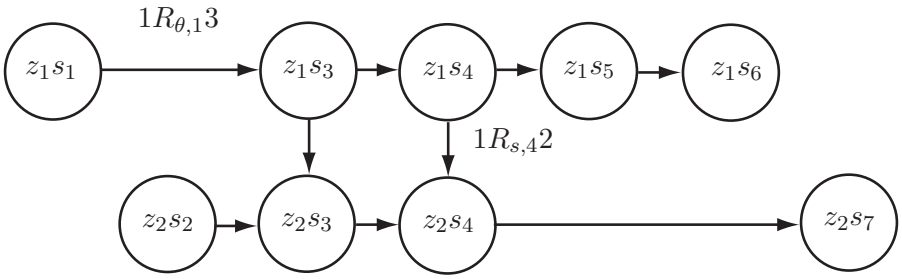


Figure 2: Event graph of a simple example system.

For each event e_{ij} are additionally defined

$$\text{maximal and minimal duration } T_{ij}^{\max} \text{and } T_{ij}^{\min} \tag{8}$$

$$\text{maximal and minimal time } t_{ij}^{\max} \text{and } t_{ij}^{\min} \tag{9}$$

$$\text{maximal and minimal velocity } v_{ij}^{\max} \text{and } v_{ij}^{\min} \tag{10}$$

$$\text{minimal buffer time } \sigma_{ij} \tag{11}$$

For a DAS a minimal velocity can be given for reasons of passenger comfort, driver acceptance or security concerns. If this is the case for an event a maximal duration is set, too. If a train's last event is followed by a stop, the intended arrival time t_{ij}^{arr} can be taken from the commercial timetable or a prediction of train running.

2.2 Computation of Drivable Train Paths with Potential for Energy-Efficient Driving for a Given Train Order

A non-linear constraint programming approach is used to determine an energy-optimal drivable paths for each train. For each event e_{ij} in R_E optimal times and velocities are searched:

$$\text{optimal time after event } e_{ij} : t_{ij}^* \tag{12}$$

$$\text{optimal velocity after event } e_{ij} : v_{ij}^* \tag{13}$$

The optimal times shall lay in intervals which can be restricted by time-minimal running for the given order of trains ($t^{\min}$) or by longest running times determined from the arrival times as given in the commercial timetable. Furthermore the minimal velocity [Alb12], if it exists, and the maximal velocity restriction of the section form the interval borders for the velocity variables. If unplanned intermediate stops are to be allowed, the maximal running times and the minimal velocities can be skipped.

$$t_{ij}^{\min} \leq t_{ij}^* \leq t_{ij}^{\max} \tag{14}$$

$$v_{ij}^{\min} \leq v_{ij}^* \leq v_{ij}^{\max} \tag{15}$$

Linear constraints are given with the minimal train separation time σ between each two consecutive train paths

$$\forall j \forall (i,l) \in R_{s,j} : t^*_{ij} + \sigma_{ij} \leq t^*_{lj} \tag{16}$$

which are like the running times T_{ij} considered as given. From T_{ij} result non linear constaints

$$\forall i \forall (j,k) \in R_{t,i} : T^{\min}_{ik}(v_{ij}, v_{ik}) \leq t^*_{ik} - t^*_{ij} \leq T^{\max}_{ik}(v_{ij}, v_{ik}). \tag{17}$$

A drivable train path is searched for each of the trains. The objective function represents the approximate energy consumption of all paths, if the time-velocity tuples act as destination points for DAS optimisation. Additionally velocities lower than the DAS advisable velocity v_{DAS} are penalised with the weighting factor ω:

$$\min \sum_{(i,j)\in R_E} \Big(\mathcal{E}_{ij}(t_{ij}, v_{ik}, v_{ij}) + \omega \max(0, v_{\mathrm{DAS}} - v_{ij})^2 \Big). \tag{18}$$

The energies $\mathcal{E}_{ij}$ summed here are the estimated minimal consumptions of DAS optimised train runs with the given start velocities v_{ij}, destination velocities v_{ik} and running times t_{ij}. For high start velocities and low destination velocities the energy consumption is lower than in other cases because of the trains high kinetic energy at the start. Lower start velocities in combination with higher destination velocities lead to higher energy consumptions. Therefore as initial guess for the optimisation process a train path with low fluctuations in velocity between train services can be used.

2.3 Computation of Envelopes out of Train Paths

To gain envelopes which offer a wide space for DAS optimisation the time intervals should be as large as possible. The following goal function is used:

$$max \sum_{(i,j)\in R_E} \Big(t^{\max *}_{ij} - t^{\min *}_{ij} \Big). \tag{19}$$

For services which are not affected by the needs of other train's services, envelopes with time intervals of maximal length are constructed by

$$((i,l) \in R_{s,j} \wedge t^{\min}_{lj} \geq t^{\max}_{ij} + \sigma_{ij}) \vee R_{s,j} = \emptyset \Rightarrow t^{max*}_{ij} = t^{\max}_{ij},\ t^{\min *}_{ij} = t^{\min}_{ij} \tag{20}$$

For all other times an optimisation using (20) as objective is performed, where for each event e_{ij} in R_E the optimal minimal and maximal times $t^{\min *}_{ij}$ and $t^{\max *}_{ij}$ after the event are searched. To ensure that the envelope built contains at least one drivable path, borders of the solution space are formed to guarantee that the computed times enclose the times t^*_{ij} of

the path computed in the previous step.

$$t_{ij}^{\min} \leq t_{ij}^{\min *} \leq t_{ij}^{*} \leq t_{ij}^{\max *} \leq t_{ij}^{\max} \tag{21}$$

The train separation times again form linear constraints to the involved time variables.

$$\forall j \forall (i,l) \in R_{s,j} : t_{ij}^{\max *} + \sigma_{ij} \leq t_{lj}^{\min *} \tag{22}$$

3 Case Study

The applicability of the algorithm to a real-world problem was verified in a case study for the single-track iron ore line in northern Sweden, which is part of the demonstrator scenarios of the ON-TIME project. A small part of the line with a length of about 50 km, maximal speeds around 60 km/h, gradients between -14 ‰ and 10 ‰ and a travel time of around one hour was chosen. A section was generated between each two signals on the line, all of the sections on the open track could be travelled in both driving directions. Two freight trains with electric locomitives and a weight of about 1500 t were simulated which crossed at an overtaking loop with one train having to stop.

To make use of the described constraint programming approaches continuous functions of possible driving times and energy consumptions needed to be computed.

3.1 Computation of Feasible Running Times

For each section the minimal and maximal running time (see formula (17)) depend on start and target velocity. These functions $t_{\min}(v_{\text{start}}, v_{\text{end}})$ and $t_{\max}(v_{\text{start}}, v_{\text{end}})$ are obtained by computing discrete velocity points of the curve by train running simulation and constructing a cubic polynomial spline from the results. The latter is used to interpolate running times in the constraint programming. An example of such a spline and its original data is given in Fig. 3. Table 1 shows mean square errors of the spline approximation of running time for different velocity pitch lengths on a sample section (minimal running time $t_{\min} = 385$ s). A spline with 7.2 km/h step-size (corresponds to 2 m/s) has finally been adopted.

pitch in $\frac{\text{km}}{\text{h}}$	mse in seconds	mse in % of $t_{\min}$
5	0.1479	0.038
7.2	0.7460	0.19
10	2.0529	0.53
15	6.3923	1.66

Table 1: Mean square error of spline approximated running times for different velocity pitches

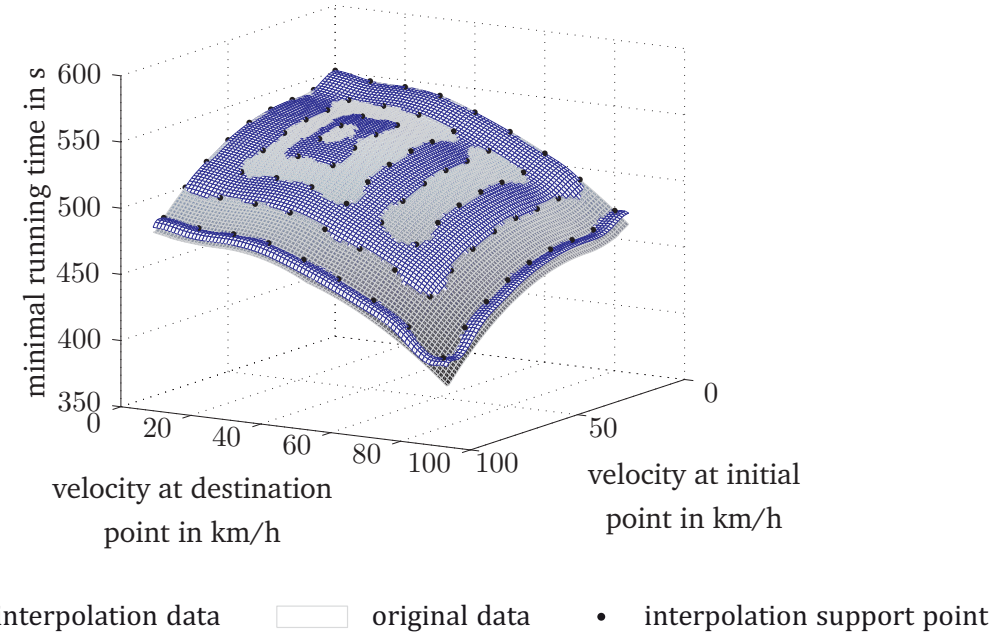


Figure 3: Minimal running time in dependence of start and target velocity; original data and cubic spline with preset border derivatives.

3.2 Energy Consumption on a Section Between Two Signals

The energy consumption of a DAS optimised run and its main influencing variables – which are namely start and target velocities, running time, track length and height profile, train parameters and the possibility to recuperate – form a hyperplane. The results obtained using train running simulation show that track length and vehicle parameters mainly affect the height of this plane while the gradient of the plane is affected by the difference of start and target velocity. It was chosen to model driving time, start velocity and target velocity the energy consumption as decision variables for drivability. So, for this case study the minimal energy consumption for each train on each section for a set combinations of start and target velocity and running time was computed using a dynamic programming approach.

It turned out that the velocities have a much higher influence on the energy consumption than the running times (see Fig. 4). With growing target velocity the energy consumption grows more than linearly. It decreases in a similar way with growing start speed. To model the dependencies between these four variables as a continuous function the MATLAB DACE toolbox was used (see [Lop02]). This toolbox provides the functionality to interpolate values out of given neighboured data samples under the assumption of stochastic dependencies by constructing a Kriging model (see also [Sas02]).

3.3 Results

The variability of the energy consumption of a single section causes the energy minimising goal function to have many local minima in which the non-linear programming solver can be

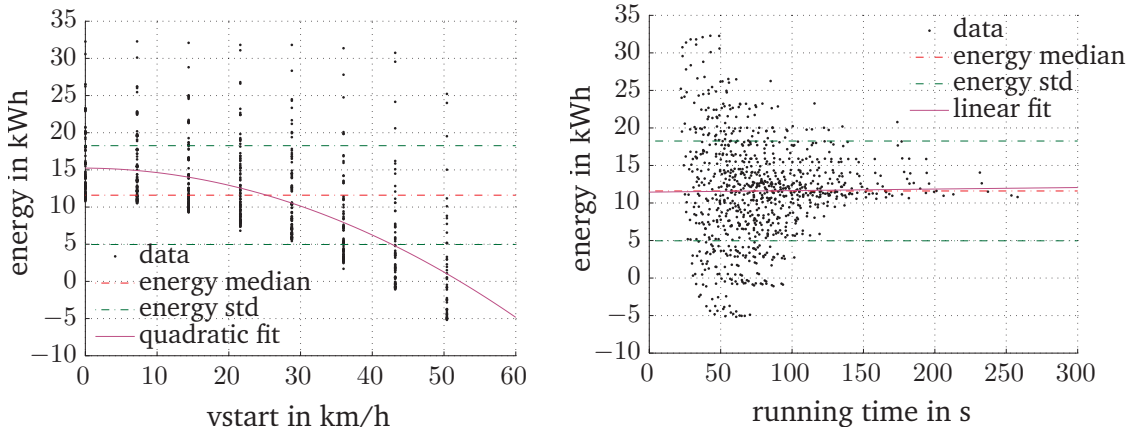


Figure 4: Energy consumption of runs with different running times and start and end target velocities shown in dependence of start velocity (left) and running time (right).

caught, particularly when used with the non-linear constraints. So, in order to get a better result, a multi-start function was used by constructing random drivable initial guesses for the optimisation process. Unfortunately minimality can still not be guaranteed.

To check the suitability of the solver for the dedicated problem, 160 runs were done with different start solutions generated with a Latin Hypercubes Designpattern on intervals of $[0, 1]$ and mapped to the drivable region for each decision variable. The obtained objective function values were rated against the best of them and their distribution obtained from all experiments is given in Fig. 5. It shows out that about 95 % of the solutions perform not worse than 104% of the best solution. The obtained velocity profiles are shown in Fig. 6. After constructing non-conflicting and reachable envelope times one gains the complete envelopes like shown in Fig. 7.

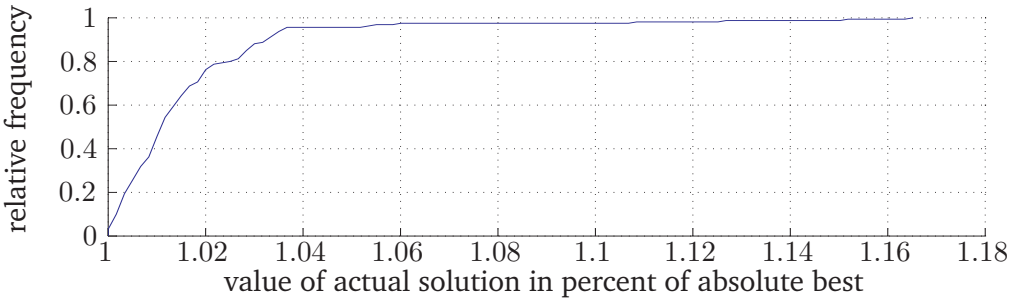


Figure 5: Distribution of goal function values of solutions computed with multistart function.

3.4 Computational Aspects

The simulation of the running times and the energy consumptions is based on train running simulation with EULER's method (described in [Jae13]). The computation of the energy consumption $\mathcal{E}_{ij}(t_{ij}, v_{ik}, v_{ij})$ consumed a noticeable amount of time (about six hours), but can be reduced by adjusting discretisation pitches in the dynamic programming approach or

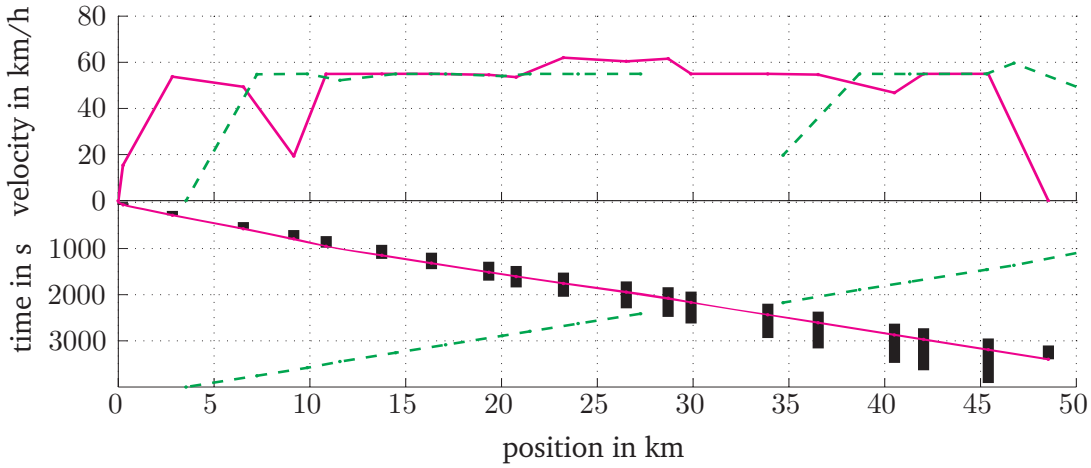


Figure 6: Energy optimised, drivable paths inside their envelopes; boxes: time windows for train one; solid/dashed lines: paths of the two regarded trains.

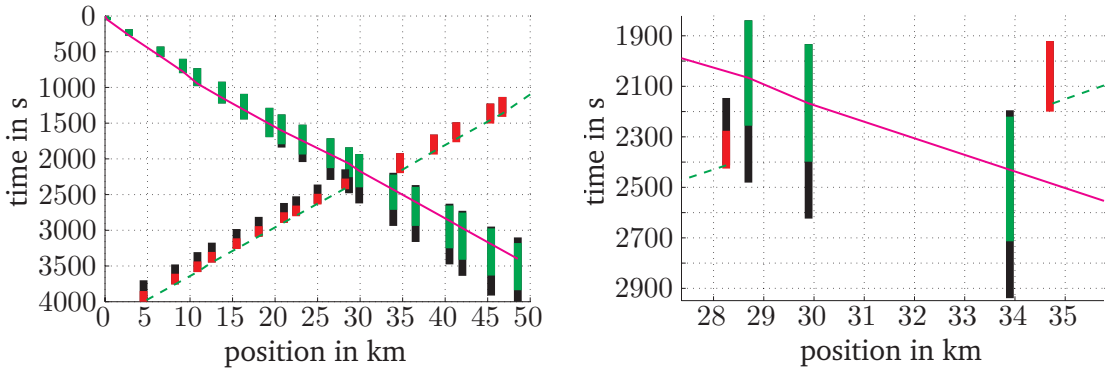


Figure 7: Optimised envelope; blocking times of train one and train two; dashed/solid: optimised paths within envelopes.

by approximating energy consumptions of different train types. These computations can be done in a pre-optimisation step with no real-time needs. The train mass is the train property with the highest influence on energy consumption, so it should be taken into account as a weighting factor. This could also reduce the amount of data needed to hold.

For the optimisation, the non-linear constraints programming solver of the MATLAB optimization toolbox was used. One optimisation run for the two train paths in the scenario takes between one and two minutes on an Intel core i5 processor with 3.1 GHz without simplification and parallelisation. It should be noted that the examined scenario is simple, but the number of constraints is quite significant with 78 decision variables with their upper and lower bounds, 156 linear constraints and nonlinear constraints (see formula (17)) for each of the variables. A significant reduction of the computation time in this scenario could be reached by inserting a simplification step before executing the optimisation, e.g. by reducing the number of sections to the overtaking loops. In this case study the number of sections would be reduced from 39 to 6 with a corresponding reduction of variables and constraints.

It is expected that through such a simplification step even complicated scenarios involving multiple trains could be formulated with less constraints and thus be computed in acceptable computation time.

4 Conclusions and Outlook

The case study shows that applying non-linear constraint programming for train path envelope construction is possible and delivers reasonable results. In order to achieve the necessary performance for an integration in real-time traffic management systems, further investigations concerning the calculation and representation of energy consumptions are needed. It has to be evolved whether a simple approximation using polynomials or splines instead of the Kriging-model would be sufficient.

Furthermore the multistart procedure is very time consuming, but could be bypassed by using a good heuristic initial guess. Also the number of starts which are needed to find a satisfying solution has to be investigated further. To get authoritative results more and more complex case studies have to be done.

Acknowledgement

This work is part of the project ON-TIME (Optimal Networks for Train Integration Management in Europe) which is co-funded by the European Commission within the Seventh Framework Programme (2007-2013), Grant Agreement FP7-SCP01-GA-2011-285243.

References

[Alb12] T. ALBRECHT, A. BINDER, and C. GASSEL: “Applications of real-time speed control in rail-bound public transtransport systems”. In: *IET Intelligent Transportation Systems* (2012). DOI: 10.1049/iet-its.2011.0187.

[DAr08] A. D’ARIANO: “Improving real-time train dispatching: models, algorithms and applications”. PhD thesis. Delft: TRAIL Research School, Apr. 7, 2008.

[Jae13] B. JAEKEL, T. ALBRECHT, and F. THONIG: “Comparative Analysis of Algorithms for Train Running Simulation”. In: *5th International Seminar on Railway Operations Modelling and Analysis RailCopenhagen2013*. IAROR. Copenhagen, May 13–15, 2013.

[Lag11] M. LAGOS: “CATO offers energy savings”. In: *Railway Gazette International* 167.5 (May 2011), pp. 50–52. ISSN: 0373-5346.

[Lop02] S. N. LOPHAVEN and et AL.: *DACE – A MATLAB Kriging Toolbox – Version 2.0*. 2002. DOI: 10.1.1.73.5824.

[Mit09] I. MITCHELL: "The Sustainable Railway: Use of Advisory Systems for Energy Savings". In: *IRSE News* 151 (2009), pp. 2–7.

[Mon09] M. MONTIGEL: "Operations control system in the Lötschberg Base Tunnel". In: *RTR - European Rail Technology Review* 2 (2009), pp. 43–44.

[OTP13] *ON-TIME Project Website*. (Last Access: 30 June 2013). URL: http://www.ontime-project.eu/.

[Pud11] P. PUDNEY, P. HOWLETT, A. R. ALBRECHT, D. COLEMAN, X. VU, and J. KOELEWIJN: "Optimal Driving Strategies with Intermediate Timing Points". In: *Proceedings of the 10th International Heavy Haul Association Conference*. Calgary, 2011.

[Sas02] M. J. SASENA: "Flexibility and Efficiency Enhancements for Constrained Global Design Optimization with Kriging Approximations". PhD thesis. University of Michigan, 2002. URL: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=68A8C2250FEF6661D1FAC7A3C759147E?doi=10.1.1.2.4697&rep=rep1&type=pdf.

[Sob11] S. SOBIERAJ, J. RODRIGUEZ, and G. MARLIERE: "Simulation of solution of a fixed-speed model for the real-time railway traffic optimisation problem". In: *4th International Seminar on Railway Operations Modelling and Analysis RailRome*. IAROR. Rome, 2011.

Corresponding author: Birgit Jaekel, Technische Universität Dresden, "Friedrich List" Faculty of Transportation and Traffic Sciences, 01062 Dresden, Germany, phone: +49 351 463 36786, e-mail: birgit.jaekel@tu-dresden.de

Kronecker Algebra based Modelling of Railway Operation

Mark Volcic, Johann Blieberger, Andreas Schöbel
Vienna University of Technology

Abstract

We present a method for modelling a railway system with serveral trains, their routes, and track sections. Our model is based on Kronecker Algebra and enables us to get all possible train movements within the system. Furthermore we can find deadlocks and calculate the travel time for each train. Blocking time is incorporated into the calculated travel time. Kronecker Algebra consists of Kronecker product and Kronecker sum and is used to manipulate matrices, which represent the train routes and the shared resource (e.g. track section).

Keywords: Kronecker Algebra, Travel Time Analysis, Deadlock Analysis, Semaphore

1 Introduction

We present a graph-based method for calculating the travel time of trains within a railway network on a fine-grained level. The routes of trains are modelled by graphs. Train sections may be part of routes of several trains. We assume that at the same time only one single train occupies a train section. Our model employs semaphores in the sense of computer science to guarantee that only one train enters a train section. These semaphores are modelled by simple graphs.

Graphs can be represented by matrices. Interestingly, simple matrix operations can be used to model concurrency and synchronization via semaphores [Mit11; Mit12b; Vol12]. These matrix operations are known as Kronecker sum and Kronecker product. With help of these operations we build a graph describing the overall railway system.

By traversing the resulting graph we compute the travel time of the trains within the network. Blocking among trains occurs due to sharing of train sections, connections, and overtaking. Blocking time is incorporated into the calculated travel time.

In comparison to simulation tools we determine all possible train movements within the railway system, we find blocking situations and calculate the travel time for each train at once. Thus our approach does not need several simulation runs to find all possible solutions (including blocking situations). Common simulation tools calculate only a single result in

one simulation run and so they can not ensure the detection of blocking situations, which might be present in the given railway system.

In [Cui10] Banker's algorithm is modified such that it can be employed for deadlock analysis in railway systems. Since Banker's algorithm has been designed for standard computer systems it is not well-suited for railway systems. For example it may prohibit allocating a resource (track section) although a potential deadlock can be bypassed. In contrast to our approach both track sections and switches have to be modelled.

In [Pac93] Movement Consequence Analysis (MCA) and Dynamic Route Reservation (DRR) are introduced for deadlock analysis. Both are rule-based methods for which correctness cannot be proved. It delivers false positives.

2 Preliminaries

This paper is based on [Mit12a], which is concerned with the Timing Analysis of Concurrent Programs. Instead of using processors or threads, we will discuss operations with trains and instead of shared memory, track sections are used for our purpose. For the synchronization of the trains on a track section semaphores in the sense of computer science (cf. [Dij65]) are used. Thus a train can enter or reserve a track section only if it is not blocked by another train using the semaphore of the track section. Blocking may occur only in succession of semaphore calls.

To model the movements of trains in a railroad system we use graphs, which are represented by adjacency matrices. We assume that the edges in a graph are labeled by elements of a semiring. Definitions and properties of the semiring can be found in [Kui86; Mit11]. Our semiring consists of a set of labels $\mathcal{L}$ representing semaphore calls denoted by p_i and v_i (cf. [Dij65]):

- p_i means reserving or entering a shared resource (e.g. track section).
- v_i means releasing or leaving a shared resource.

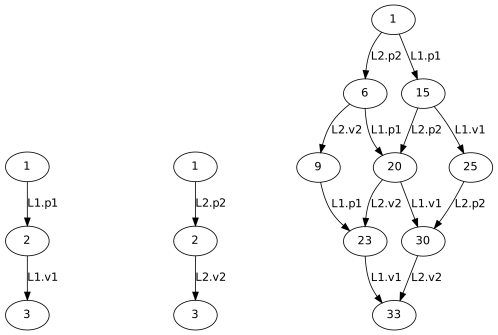
Usually two or more distinct train route graphs refer to the same track section to model synchronization.

2.1 Modeling Synchronization and Interleavings

Kronecker product and Kronecker sum form Kronecker algebra. In the following we define both operations. From now on we use matrices out of $\mathcal{M} = \{M = (m_{i,j}) | m_{i,j} \in \mathcal{L}\}$ only.

Definition 1 (Kronecker product) *Given a m-by-n matrix A and a p-by-q matrix B, their* Kronecker product *denoted by $A \otimes B$ is a mp-by-nq block matrix defined by*

$$A \otimes B = \begin{pmatrix} a_{1,1} \cdot B & \cdots & a_{1,n} \cdot B \\ \vdots & \ddots & \dots \\ a_{m,1} \cdot B & \cdots & a_{m,n} \cdot B \end{pmatrix}$$



(a) Train L_1 **(b)** Train L_2 **(c)** $L_1 \oplus L_2$

Figure 1: Example: Different track sections

Kronecker product allows to model synchronization (cf. [Buc02; Mit11; Pla85]).

Definition 2 (Kronecker sum) *Given a matrix A of order m and a matrix B of order n, their* Kronecker sum *denoted by $A \oplus B$ is a matrix of order mn defined by $A \oplus B = A \otimes I_n + I_m \otimes B$, where I_m and I_n denote identity matrices of order m and n respectively.*

In general, a railroad system consists of a finite number of trains and track sections which are represented by graphs. As already mentioned the graphs are represented by adjacency matrices, the track sections are represented by binary semaphores in the sense of computer science. The matrices have entries which are referred to as labels $l \in \mathcal{L}$.
Formally, the system model consists of the tuple $\langle \mathcal{T}, \mathcal{S}, \mathcal{L} \rangle$, where $\mathcal{T}$ is the set of graph adjacency matrices describing routes of trains, $\mathcal{S}$ refers to the set of adjacency matrices describing track sections (semaphores) and the labels in $T \in \mathcal{T}$ and $S \in \mathcal{S}$ are elements of $\mathcal{L}$, respectively. The matrices are manipulated by using Kronecker algebra.

As already mentioned, there is a correspondence between matrices and graphs. In general a directed labelled graph $C = \langle V, E, n_e \rangle$ consists of a set of labelled nodes V, a set of labelled edges $E \subseteq V \times V$ and an entry node $n_e \in V$. The sets V and E are constructed out of the elements of $\langle \mathcal{T}, \mathcal{S}, \mathcal{L} \rangle$. In this paper we label graph nodes simply by positive integers. Correspondence between graphs and matrices – frequently called *adjacency matrices* – is as follows: If there exists an edge labelled a from node i to node j, then the corresponding adjacency matrix M has $m_{i,j} = a$. If there is no edge between node i and node j, then $m_{i,j} = 0$. The following example illustrates some interleavings of two train routes in a railway system and how Kronecker sum handles it.

2.2 Example: Different track sections

Let the matrices $L_1 = \begin{pmatrix} 0 & p_1 & 0 \\ 0 & 0 & v_1 \\ 0 & 0 & 0 \end{pmatrix}$ and $L_2 = \begin{pmatrix} 0 & p_2 & 0 \\ 0 & 0 & v_2 \\ 0 & 0 & 0 \end{pmatrix}$. The graphs of matrices L_1 and L_2 are shown in Fig. 1a and Fig. 1b, respectively. Now interpret L_1 and L_2 as being train

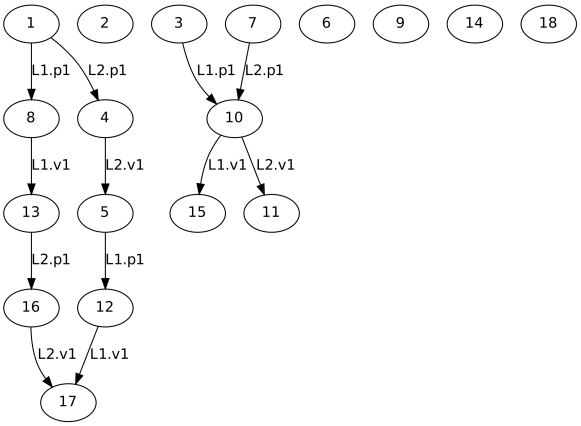


Figure 2: Graph of $R = (L_1 \oplus L_2) \otimes S$

routes and p_1, v_1, p_2, and v_2 as being actions of the trains with the following meaning: p_1 denotes train L_1 enters track section 1, v_1 means L_1 has left track section 1, p_2 denotes train L_2 enters track section 2, and v_1 means train L_2 has left track section 2. All possible timing interleavings by executing L_1 and L_2 are shown in Fig. 1c.

2.3 Example: Same track section

Now assume that both trains use the same track section. It is clear that in this case the temporal interleavings of the previous example are no more valid. The trains have to synchronize in order to perform their actions correctly. This can be modelled by Kronecker product and an additional matrix of the form $S = \begin{pmatrix} 0 & p \\ v & 0 \end{pmatrix}$ where p denotes the action *Enter the track section* and v means *Train has left the track section*. The correct system behavior can be described by the matrix $R = (L_1 \oplus L_2) \otimes S$. As a result the graph will be decomposed into sub-graphs (Fig. 2). Clearly only the part reachable from the entry node (node 1) is responsible for the system behavior, the others can safely be ignored. Since node 1 is the entry node, we see that now the trains enter the track section one after the other. Note that the two paths in the subgraph correctly mirror the two cases where L_1 enters the track section before L_2 and vice versa. A proof that Kronecker product models synchronization correctly can be found in [Mit11]. To increase readability of the resulting graph we distinguish the following node types:

- Red nodes denote deadlocks[1] or nodes from which only deadlocks can be reached (The final node can not be reached from such nodes).

- Green nodes denote *safe* states. A state is *safe* if all trains can perform their actions without having to take into account the moves of the other trains in the system, provided that the track section which they are to enter is not occupied by another train[2].

[1] Deadlock analysis for railway systems via our approach is studied in [Mit12b].

[2] If a track section is occupied by another train, the movement of the train wanting to enter may be delayed (blocked) but no deadlock can occur.

- From orange nodes both red and green nodes can be reached.
- Filled nodes are *synchronizing* nodes (at least two trains must be synchronized).

3 System Model

We model a general railway system S by a set of track sections $T = \{T_i | 1 \leq i \leq r\}$. Each track section T_i is modelled by matrix $T_i = \begin{pmatrix} 0 & p_i \\ v_i & 0 \end{pmatrix}$. In addition, a railway system consists of a set of trains $L = \{L_j | 1 \leq j \leq t\}$. The route R_j of train L_j is a sequence of track sections $T_{l_1}, \ldots, T_{l_s}$ for $1 \leq l_n \leq r$ and $1 \leq n \leq s$. Each route is modelled by a $2s \times 2s$-matrix. The set of routes is denoted by $R = \{R_j | 1 \leq j \leq t\}$.
The behavior of railway systems $S\langle T, L, R\rangle$ is modelled by

$$S = \left(\bigoplus_{j=1}^{t} R_j \right) \otimes \left(\bigoplus_{j=1}^{r} T_i \right)$$

where during the evaluation of the Kronecker product, for simplicity we let $p_i = p_i \cdot p_i$ and $v_i = v_i \cdot v_i$.
The different paths in the graph corresponding to matrix S mirror all possible behaviors of the railway system in terms of temporal interleavings of the actions of trains, namely entering and leaving track sections.
Special cases like overtaking or waiting for other trains, which need some additional semaphores are discussed in the following examples.

4 Travel Time Analysis

Each node of the graph is assigned a variable and an equation is setup based on the predecessors of the node. A variable is represented by a vector, where each component of the vector corresponds to a train. The equations are used to calculate the worst-case travel time for each train.
Synchronizing nodes are nodes where blocking occurs. These nodes have an incoming edge labeled by a semaphore v-operation, an outgoing edge labeled by a p-operation of the same semaphore, and these edges are part of different trains. In this case the train with the p-operation has to wait until the other train's v-operation is finished.

Let the vector $\mathfrak{X} = (X_1, \ldots, X_l, \ldots, Xp)^T$. We write $\mathfrak{X}^{(l)} = X_l$ to denote the lth component of vector $\mathfrak{X}$.

Definition 3 *Let* $\mathfrak{X} = (X_1, \ldots, X_p)^T$ *and* $\mathfrak{Y} = (Y_1, \ldots, Y_p)^T$. *Then we define*

$$\max(\mathfrak{X}, \mathfrak{Y}) := (\max(X_1, Y_1), \ldots, \max(X_p, Y_p))^T.$$

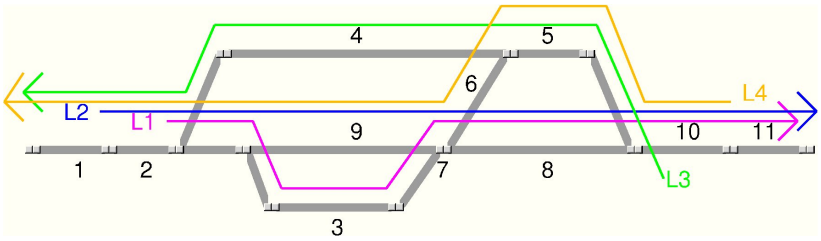


Figure 3: Railway system

Definition 4 *A* synchronizing node *is a node s such that*

- *there exists an edge $e_{in} = (i, s)$ with label v_k and*
- *there exists an edge $e_{out} = (s, j)$ with label p_k,*

where k denotes the same semaphore and e_{in} and e_{out} are mapped to different trains.

Definition 5 (Setting up equations) *If n is a non-synchronizing node, then*

$$\mathfrak{x}_n = \max_{k \in Pred(n)} (\mathfrak{x}_k + t(k \to n)),$$

where the lth component of vector $t(k \to n)$ is the time assigned to edge $k \to n$ and edge $k \to n$ is mapped to train l. The other components of $t(k \to n)$ are zero. The set of predecessor nodes of node n is referred to as $Pred(n)$.
Let s be a synchronizing node. In addition, let λ_i and λ_j be the trains where the edges $i \to s$ and $s \to j$ are mapped to. Then for $l \neq \lambda_j$

$$\mathfrak{x}_s^{(l)} = \max_{k \in Pred(s)} \left(\mathfrak{x}_k^{(l)} + t(k \to s)^{(l)} \right) \tag{1}$$

and

$$\mathfrak{x}_s^{(\lambda_j)} = \max \left(\mathfrak{x}_i^{(\lambda_i)} + t(i \to s)^{(\lambda_i)}, \mathfrak{x}_k^{(\lambda_j)} + t(k \to s)^{(\lambda_j)} \right) \tag{2}$$

where the first term considers the incoming v-edge and the second one the incoming edge of train λ_j.

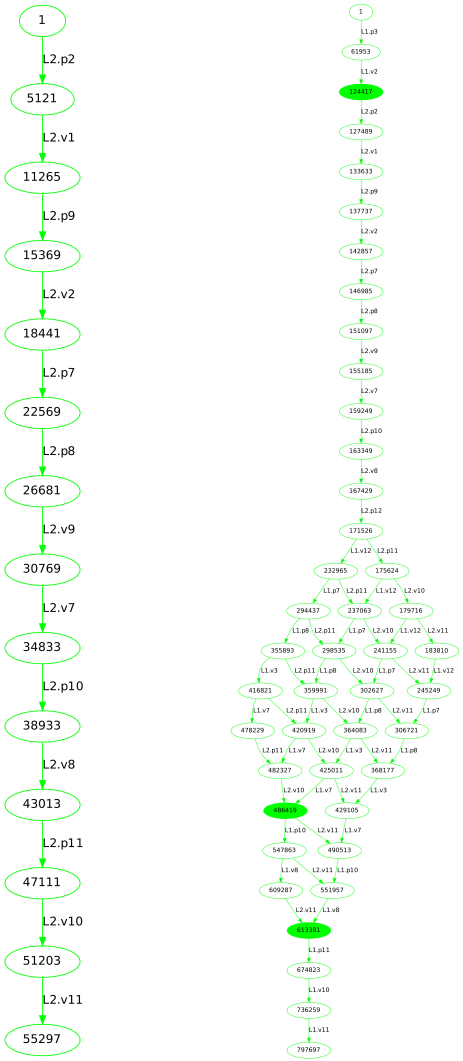
The system of equations can be solved easily by inserting one equation into another. An example of setting up the equations is be given in section 5.2. Our approach enables us to calculate the travel time (including the blocking time) for each train in a complex railway system. Blocking might occur due to occupied track sections or connections for instance.

5 Examples

In this section we give a small example which is represented in Fig. 3. At first we calculate the travel time for one train (L_2). After that we add the other trains (L_2, L_3 and L_4) step by

step and analyze the results. To calculate the travel time for the following examples we make the following assumptions:

- Trains L_1 and L_3 are slow trains and need two time units for each track section.
- Trains L_2 and L_4 are fast trains and need only one time unit for each track section.



(a) Example I **(b)** Example II

Figure 4: Results of Example I and II

5.1 Example I

Train L_2 starts at track section 1 and has the following route:

$$R_2 = p_2, v_1, p_9, v_2, p_7, p_8, v_9, v_7, p_{10}, v_8, p_{11}, v_{10}, v_{11}$$

As there is only one train within the system, the resulting graph only consists of the 13 operations along the train's route, which result in a graph with 14 nodes (Fig. 4a). The

resulting travel time is the sum of the travel time of each track section and is calculated to be 6 time units.

5.2 Example II

Now the train L_1 is added to our railway system, which starts at track section 2. At section 3 it has to stall until it is overtaken by train L_2. As a result there must be a synchronization between the two trains and thus the travel time of train L_2 increases due to the blocking time. To ensure synchronization of two trains an (additional) *artificial track section* is added to the routes (track section 12). These track sections are used to ensure the desired order of some trains. The modified routes of the two trains read as follows:

$$\begin{aligned} R_1 &= p_3, v_2, v_{12}, p_7, p_8, v_3, v_7, p_{10}, v_8, p_{11}, v_{10}, v_{11} \\ R_2 &= p_2, v_1, p_9, v_2, p_7, p_8, v_9, v_7, p_{10}, v_8, p_{12}, p_{11}, v_{10}, v_{11} \end{aligned}$$

The resulting graph has 44 nodes, including 3 synchronizing nodes and can be found in Fig 4b. The travel time of train L_1 is calculated to be 13 time units and that of train L_2 is 8 time units because train L_2 has to wait caused by a headway conflict and train L_1 has to wait until the overtaking has been finished.

Calculation of the travel time for non-synchronizing nodes

To calculate the travel time for a non-synchronizing node node (e.g. $\mathfrak{x}_{241155}$), the travel time of each predecessor plus the travel time of the connecting-edge are computed. So we get the travel time for each incoming edge. The maximum of these values is taken as the travel time for node $\mathfrak{x}_{241155}$. For example node $\mathfrak{x}_{237063}$ is connected to $\mathfrak{x}_{241155}$ by an edge labeled with $L_2.v_{10}$, which means that train L_2 executes v_{10}. As a result v_{10} is used in the second vector of the equation for L_2. Because there are no other trains involved in the transition from $\mathfrak{x}_{237063}$ to $\mathfrak{x}_{241155}$, the other value of the vector is set to 0. The same procedure is done for node $\mathfrak{x}_{179716}$, which is connected to $\mathfrak{x}_{241155}$ by an edge labeled with $L_1.v_{12}$ and thus the vector will have value 0 for the second train, and v_{12} for the first one. The complete equation for node $\mathfrak{x}_{241155}$ reads as follows:

$$\mathfrak{x}_{241155} = \max\left(\mathfrak{x}_{237063} + \begin{pmatrix} 0 \\ v_{10} \end{pmatrix}, \mathfrak{x}_{179716} + \begin{pmatrix} v_{12} \\ 0 \end{pmatrix}\right)$$

Calculation of the travel time for synchronizing nodes

As described in equation (1) and (2), there is a difference between trains which should be synchronized with others and trains which act independently (at the current node). For independent trains, the equation is setup as for non-synchronizing nodes (second line in the vector for node $\mathfrak{x}_{486419}$). The first line, which describes the travel time for train L_1 will use the travel time value of the second train L_2 because L_1 has to be synchronized with L_2 and

thus train L_1 has to wait for L_2. If there would be other incoming edges with an action from train L_1, then their travel time values would also be considered in the equation and the maximum of these values would be taken as the train's travel time at the current node. The complete equation for node $\mathfrak{x}_{486419}$ reads as follows:

$$\mathfrak{x}_{486419} = \begin{pmatrix} \mathfrak{x}^{(2)}_{482327} + v_{10} \\ \max\left(\mathfrak{x}^{(2)}_{482327} + v_{10}, \mathfrak{x}^{(2)}_{425011}\right) \end{pmatrix}$$

5.3 Example III

The last example contains the railway station with all four trains, which results in two situations where trains must be synchronized:

- Train L_2 has to overtake train L_1 (as in the example above)
- Train L_4 has to overtake train L_3

Thus, there must be an additional artificial track section with p_{13} and v_{13}. The routes of the four trains read as follows:

$$\begin{aligned} R_1 &= p_3, v_2, v_{12}, p_7, p_8, v_3, v_7, p_{10}, v_8, p_{11}, v_{10}, v_{11} \\ R_2 &= p_2, v_1, p_9, v_2, p_7, p_8, v_9, v_7, p_{10}, v_8, p_{12}, p_{11}, v_{10}, v_{11} \\ R_3 &= p_5, v_{10}, p_4, v_5, v_{13}, p_2, v_4, p_1, v_2, v_1 \\ R_4 &= p_{10}, v_{11}, p_5, v_{10}, p_6, v_5, p_7, p_9, v_6, v_7, p_2, v_9, p_{13}, p_1, v_2, v_1 \end{aligned}$$

The resulting graph has 992 nodes and thus 992 equations must be solved to calculate the traval time of each train. Our implementation of the algorithm will calculate the travel time for this example within 400 ms. A detailed description of the algorithm can be found in [Vol12].

The resulting graph of this example can be be found in Fig. 6 (including deadlocks). Due to the fact that we are only interested in deadlock-free situation, a reduced graph can be generated which contains all train movements which don't result in a deadlock (Fig. 5).

The results of these three examples can be found in Table 1.

Example	Number of nodes	Travel times			
		L_1	L_2	L_3	L_4
I	14		6		
II	44	13	8		
III	992	15	15	8	10

Table 1: Results of the examples

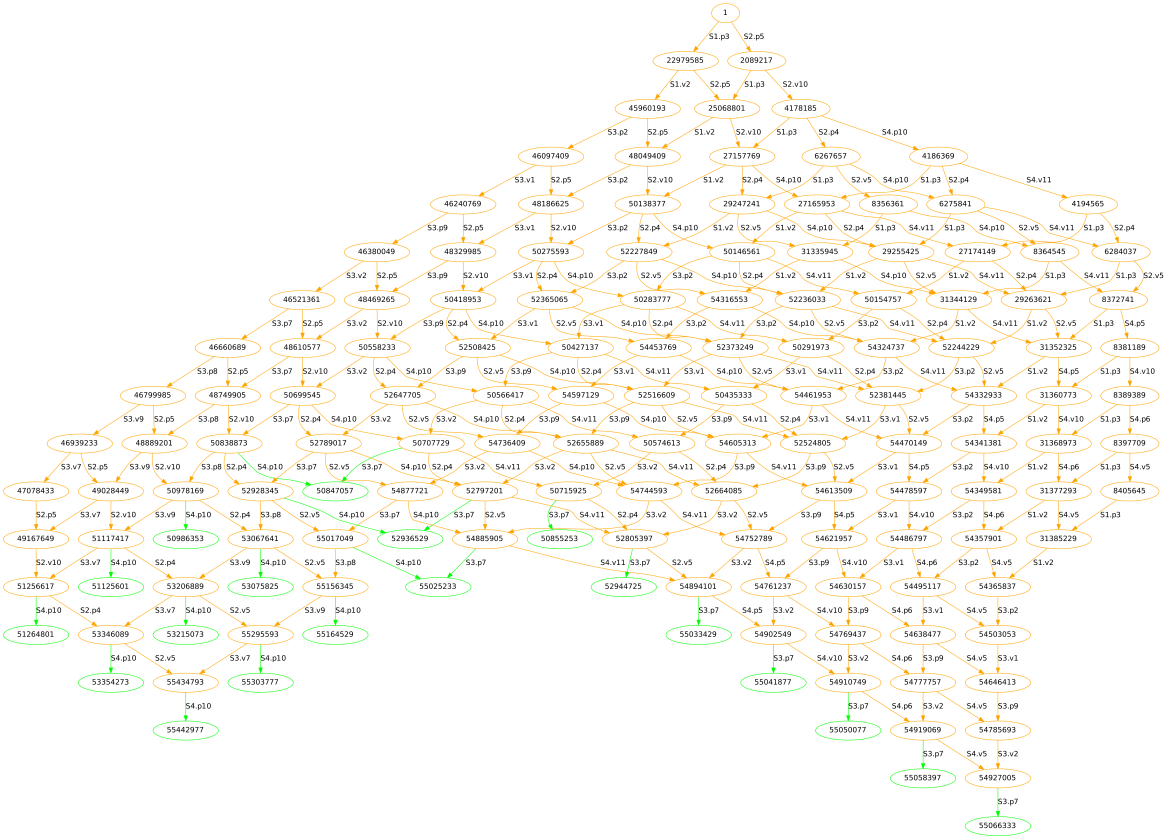


Figure 5: Reduced graph of Example III

6 Conclusion

We have presented a graph-based method for calculating travel time of trains and finding deadlocks within a railway network on a fine-grained level. Kronecker algebra is applied to manipulate matrices and to create graphs, which are represented by adjacency matrices. Our approach can be used to model complex railway systems including the calculation of travel times and the aspects of being deadlock-free and being conflict-free.

Acknowledgements

The authors would like to thank the *Austrian Ministry of Transport, Innovation and Technology* (bmvit) and the *Austrian Research Promotion Agency* (FFG) for sponsoring the project *EcoRailNet* in the frame of the program *New Energy 2020* (Project-ID: 834586).

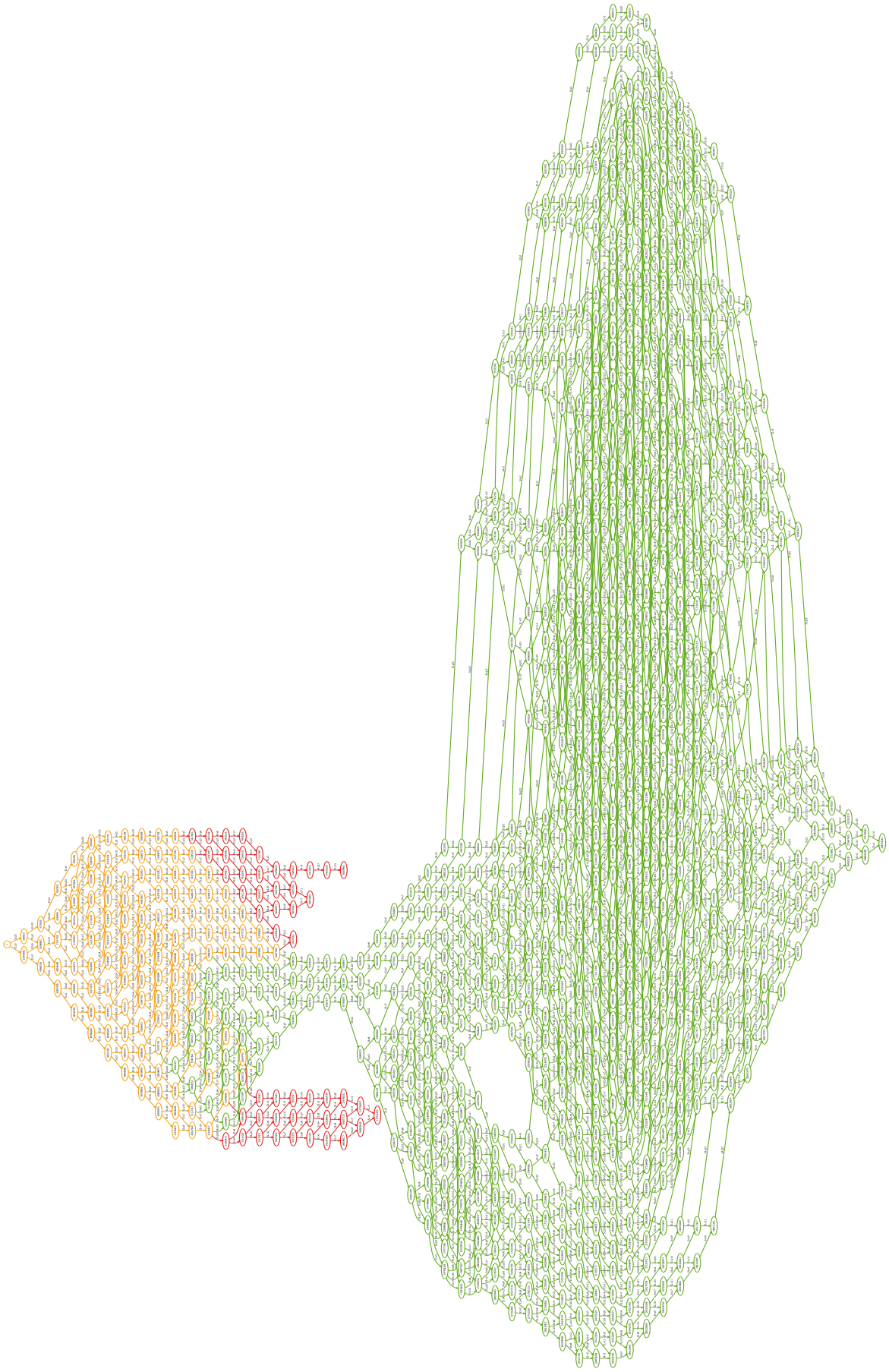


Figure 6: Resulting graph of Example III

References

[Buc02] P. BUCHHOLZ and P. KEMPER: "Efficient Computation and Representation of Large Reachability Sets for Composed Automata". In: *Discrete Event Dyn. Systems* 12.3 (2002), pp. 265–286.

[Cui10] Y. CUI: "Simulation-Based Hybrid Model for a Partially-Automatic Dispatching of Railway Operation". PhD thesis. Universität Stuttgart, 2010.

[Dij65] E. W. DIJKSTRA: "Over Seinpalen". 1965. URL: http://www.cs.utexas.edu/users/EWD/ewd00xx/EWD74.PDF.

[Kui86] W. KUICH and A. SALOMAA: *Semirings, Automata, Languages*. Springer, 1986.

[Mit11] R. MITTERMAYR and J. BLIEBERGER: *Shared Memory Concurrent System Verification using Kronecker Algebra*. Tech. rep. 183/1-155. Automation Systems Group, TU Vienna, Sept. 2011. URL: http://arxiv.org/abs/1109.5522.

[Mit12a] R. MITTERMAYR and J. BLIEBERGER: "Timing Analysis of Concurrent Programs". In: *Proc. 12th International Workshop on Worst-Case Execution Time Analysis*. Pisa, Italy, July 2012.

[Mit12b] R. MITTERMAYR, J. BLIEBERGER, and A. SCHÖBEL: "Kronecker Algebra based Deadlock Analysis for Railway Systems". In: *Promet-Traffic & Transportation* 5 (2012), pp. 359–369. ISSN: 1848-4069.

[Pac93] J. PACHL: "Steuerlogik für Zuglenkanlagen zum Einsatz unter stochastischen Betriebsbedingungen". PhD thesis. TU Braunschweig, 1993.

[Pla85] B. PLATEAU: "On the Stochastic Structure of Parallelism and Synchronization Models for Distributed Algorithms". In: *ACM Sigmetrics*. Vol. 13. 1985, pp. 147–154.

[Vol12] M. VOLCIC, J. BLIEBERGER, and A. SCHÖBEL: "Kronecker Algebra based Travel Time Analysis for Railway Systems". In: *FORMS/FORMAT 2012 – 9th Symposium on Formal Methods for Automation and Safety in Railway and Automotive Systems*. Braunschweig, Germany, Dec. 2012, pp. 273–281. ISBN: 978-3-9803363-3-8.

Corresponding author: Mark Volcic, Vienna University of Technology, Institute of Computer Aided Automation, 1040 Vienna, Austria, e-mail: mvolcic@auto.tuwien.ac.at, phone: +43 1 58801 18313

Role of Systems Engineering in Evaluation of ITS Systems – Example of the Train Dispatcher System (in Poland)

Grzegorz Karon[1], Jerzy Mikulski[2]

[1] Silesian University of Technology

[2] PSTT Polish Association of Transport Telematics

Abstract

Systems Engineering integrates all the disciplines and specialty groups into a team effort forming a structured development process that proceeds from concept to production to operation. In 2010 a questionnaire study was performed among the users of railway ITS subsystems and the survey results allowed formulating the proposals of changes to the functional structure of the systems. Evaluation of this systems, based on systems engineering (V-Model) and results of questionnaire survey (as one of others validation procedure), have been presented.

Keywords: Systems Engineering, ITS, Train Dispatcher Support System, System of Operational Work Evidence, System of Temporary Warnings Registration, Validation, Verification, Survey Methods in Transport.

1 Introduction

Train dispatcher is a highly skilled job position directly involved in the train operations within the relevant signalling control areas and on the adjacent routes or railway sections. *Line controller* is a member of the current supervision staff responsible for regulation and coordination of the whole of the activities related to the train traffic in a specific railway area, indirectly, by means of the orders issued to the train dispatchers. Similarly as in the case of a line controller, the dispatcher's job requires the knowledge of a number of instructions and regulations, however it is this particular position which bears a direct responsibility for the safe and regular train operations. The duties of the train dispatcher include making decisions about the correct preparation of the route, about the right sequence and direction of train dispatching, in agreement with the current rules and with the timetable. Above all, the duty of the train dispatcher is to react immediately in case of a danger or a disturbance of a normal train operation due to the emergency situations and to the deviations from the timetable operations. The scope of duties of a train dispatcher includes also the registration

and issuing of the *current warnings of speed limit*. The *current warnings* are all speed limitations not included in the *Permanent Speed Limitations Register* and which have been introduced by the current orders. These warnings are registered in the *Current Warning Log*. The SERWO system (*System of Electronic Propagation of Current Warnings*) has to support the correct and efficient implementation of these processes. This is an application supporting the train dispatchers in the area of registering, issuing and handling of current warnings. A part of the system is a database storing the information about the railway lines, the routes, the trains and the reasons for the warnings. It allows (among others) the printout of the orders and stores an electronic log of current warnings. It substitutes the old wire message system for sending the information on the warnings and for confirming the receipt and the recording of a warning [Her13, Kar12a]. The train dispatcher is supported in the decision making process by the SWDR system (*Train Dispatcher Support System*). This system contains all the information needed by the train dispatcher such as the timetables, planned train runs, train delays (and the reasons for these), planned and actual parameters of the trains, the trains carrying hazardous materials and trains with oversize loading gauge as well as the routes of all trains. The SEPE (*System of Operational Work Evidence*) is another system which, apart from the information about train operation, stores the information about all the disturbances to the train runs. Currently the PLK (Polish Railway Lines) makes the infrastructure available to over 50 carriers. The data of the trains which are planned to be running are input by the planning controllers into the SEPE system. The system displays the actual hour of the departure of the train as well as the hour of its arrival at the terminal station together with the information of any delays or arrivals ahead of time, in minutes. This information allows to control in real time the execution of the operational plan and the analysis of any incurred delays. The controllers input the information about the actual run of the train (hours of calling at each subsequent station, train data such as its gross weight, length, type of engine – the so called train analysis). This data is normally received by telephone from the train dispatchers operating at individual stations of the *Area Railway Line Bureaus*. The registration of the incidents and events is performed based on the reports from train dispatchers and is being input in the system by the line controllers. The stored information contains the data on the time of the event, time at which the event ended, on the exact location of the event, its influence on the operation of the trains (i.e. the delays) as well as the person in default of the event. The SEPE system permits the line controller to record the actual graph of the train run, which includes among others: annual schedule, the individual schedule, the actual routes, line track closures, permanent and temporary warnings, arrival, departure or passing times, train delays and the reasons for these [Her13, Kar12a].

The systems described herein, supporting the work of the train dispatchers and of the line controllers are continuously developed and modified as to adapt them to the changing requirements and conditions of the job positions at which the systems are used.

The survey results described in this paper and in [Her11b, Kar12b] allowed formulating the proposals of changes to the functional structure of the systems (Fig 1). For example, version 2.2 of the system includes the module of trains ride time input (operated by line dispatchers) and a subsequent change should provide an option of inputting the train analyses (concerning the trains deployed by the dispatchers).

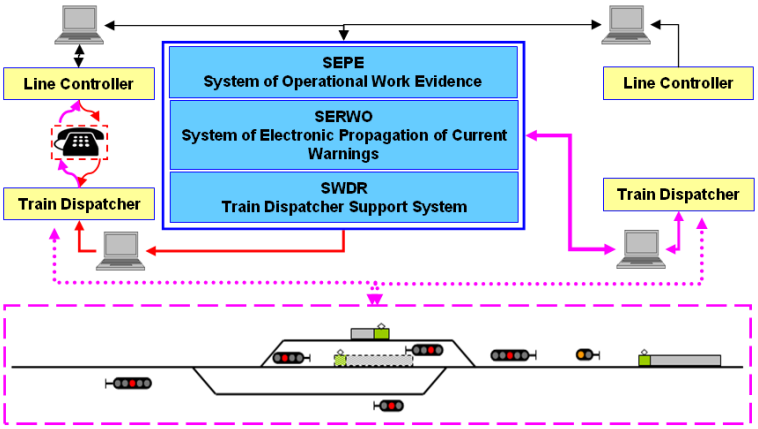


Figure 1: Communication diagram between *Train dispatcher*, *Line controller* and Supporting Systems: on the left side – without modifications, on the right side – proposed changes.

2 Verification and validation in systems engineering

As mentioned in [FHW07] definition of systems engineering by INCOES (*International Council on Systems Engineering*) is: Systems Engineering is an interdisciplinary approach and means to enable the realization of successful systems. It focuses on defining customer needs and required functionality early in the development cycle, documenting requirements, then proceeding with design synthesis and system validation while considering the complete problem. ITS projects are identified and funded through transportation planning [Kar11, Kar12d] and programming/budgeting processes in each state, planning region (e.g., metropolitan planning area) [Kar12c], and agency. The "V" diagram and the systems engineering process (Fig 2) begin once a need for an ITS project has been identified. The early steps in the "V" define the project scope and determine the feasibility and acceptability as well as the costs and benefits of the project. The latter steps support project implementation, then transition into operations and maintenance, changes and upgrades, and ultimate retirement or replacement of the system [FHW06]. Systems Engineering integrates all the disciplines and specialty groups into a team effort forming a structured development process that proceeds from concept to production to operation. In systems engineering there are distinction between verification and validation. This is an important distinction because there are lots of examples of well-engineered products that met all of their requirements but ultimately failed to serve their intended purpose. Verification confirms that a product meets its specified requirements - verification ensures that system is built right. There are four basic techniques to verify requirements: test, demonstration, inspection and analysis. Summary of the verification results should provide evidence that the system (subsystem, component) meets the requirements and identify any corrective actions that were recommended or taken as a result of the verification process . It is important that stakeholders and end users should also be materially involved in verification, particularly in the system verification activities. In systems engineering as the verification proceeds from detailed component verification to end-to-end system verification, the implementation team (software and hardware specialists) becomes less involved and the stakeholders become

more involved. The majority of system verification can be performed before the system is deployed [FHW07]. Validation confirms that the product fulfills its intended use - validation ensures that we built the right system. Validation can't be completed until the system is in its operational environment and is being used by the real users. This step in V-model is when the system has been put into operation and is beginning to be used for its intended purpose (Fig 2).

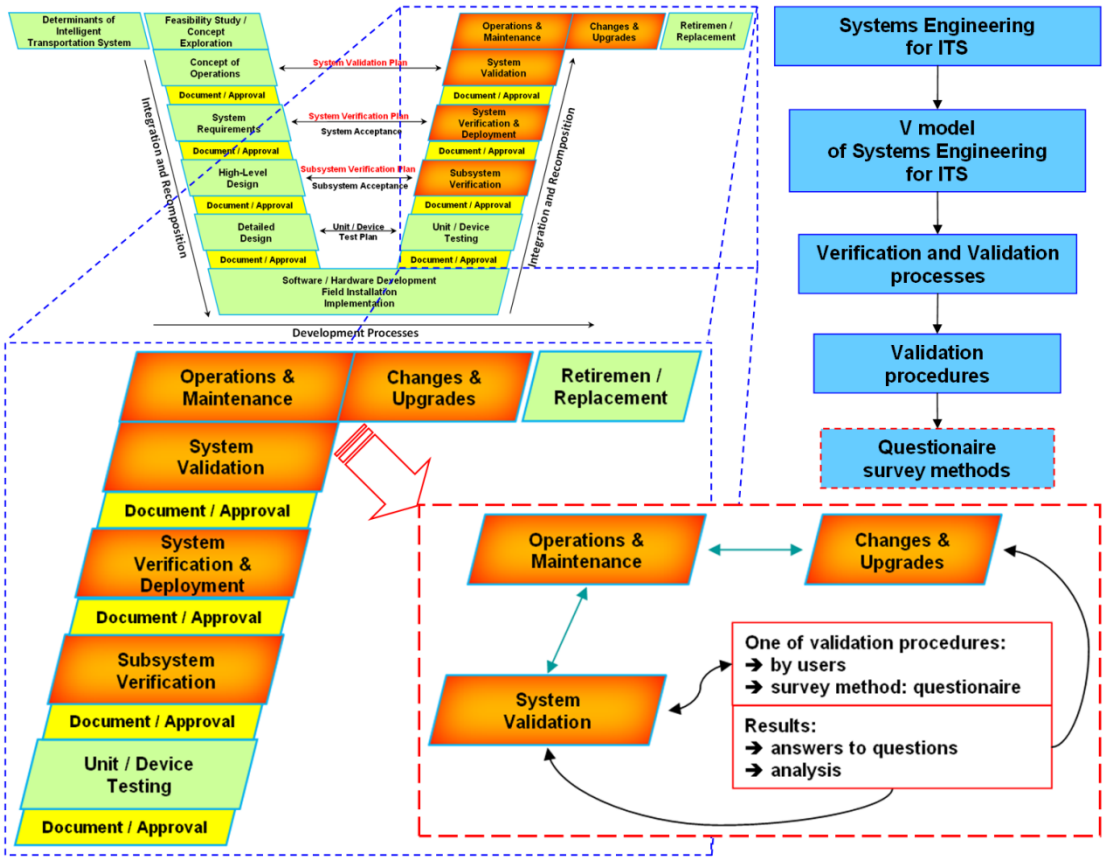


Figure 2: Proposed validation procedure as one of others in "V" diagram of systems engineering for ITS. Source: own work based on V-diagram from [FHW07].

But systems engineering approach, called *in-process validation*, seeks to validate the products that lead up to the final operational system to maximize the chances of a successful system validation at the end of the project. In-process validation is performed on an ongoing basis throughout the process by decision makers during the initial feasibility study and by stakeholders during the system concept of operations development, system requirements, system designing, building, implementation, and finally during the operation. Key aspect of validation during the implementation and operation is validating the user interface design since it has a strong influence on user satisfaction. In addition to objective performance measures (quantitative and qualitative), the system validation may also measure how satisfied the users are with the system. This can be assessed directly using surveys, interviews, in-process reviews, and direct observation [FHW07]. Between verification and validation is initial deployment. In this step system is installed in the operational environment and transferred from the project development team to the end-user that will own and operate it. The transfer also includes support equipment, documentation, operator training, and other enabling products that support ongoing system operation and

maintenance. It is important that all operations and maintenance staff should be in place and properly trained [FHW07]. During this period, operators, maintainers, and users of the system may identify issues, suggest enhancements, or identify potential efficiencies. The system will evolve over its lifetime as stakeholder priorities change and technology advances. Changes can also result from user-reported issues and recommendations and from system improvements identified from the review of operational data [FHW07]. Proposed validation procedure based on questionnaire survey among end-users (example in next topic) is shown in Fig. 2 as part of validation process in V-model of systems engineering.

3 Example of functionality evaluation of the train dispatcher system by users – selected problems

Effective and quick data input and acquisition during a railway operations and the proper processing of the data by the users (train dispatchers and traffic controllers) are the main objectives of the systems supporting the railway staff. A survey performed in 2010 allowed a functional evaluation of the SWDR (*Train Dispatcher Support System*) as well as the collection of information on suggested future modifications of the system (Fig 1). Modifications was proposed by the authors of surveys based on the analysis of respondents' answers to the open questions. The module of train running time input operated by line dispatchers is primary modification that was introduced after survey. Next modification should be an option of inputting the train analyses concerning the trains departures by the dispatchers [Her13, Kar12b]. Aim of this surveys was to evaluate the system's functionality by users in terms of individual assessment (perception) of these systems and usefulness of the information in operational work (daily operational). Next step of system evaluation will be statistical evaluation of train punctuality for both the periods before and after the installation of the systems.

The surveys were carried out mainly using the *CAWI* method (*Computer Aided Web Interview* - a questionnaire available on a Web page) - 150 questionnaires (42.1%). In order to include the persons not using the Internet (or using it incidentally) in the survey, the study was performed in parallel using *CAPI* (*Computer Aided Personal Interview*) - 102 questionnaires (28.7%), *CATI* (*Computer Aided Telephone Interview*) and *PAPI* (*Paper and Pencil Interview*) methods as well as by regular mail - 104 questionnaires (29.2 %). Taking into account the fact that in the scale of the whole railway network of PLK (Polish Railway Lines) the SEPE and SWDR systems was (at the moment when the survey was carried out) installed at 1696 work posts and with an average of 5 persons of staff operating at one work post, the estimated total population consists of 8480 persons having contact with the systems evaluated. Therefore, the collected sample of 356 questionnaires (with incomplete or erroneous ones rejected) constituting 4.2% of the population may be considered a representative one (the estimated maximum error of the survey is d=5%).

Almost half of the respondents (49%) have noticed an improvement in the punctuality of the trains as an effect of the implementation of SWDR and SEPE systems (Fig 3a). However, at the same time a large percentage of respondents (39%) see no relation between the functionality of the systems and the punctuality of the trains. Such evaluation may come from

the train dispatchers of the station with small volume of rail traffic, small number of operations and small number of delayed trains or from rail traffic small stations (points) with simple layout and one railway junction which can not change the order of trains. A second reason for the latter answers may be low awareness of the users as to the right utilization of the information from the system in further traffic management (this has also been confirmed by the fact that 12% of responses were 'Do not know') [Her13, Kar12b].

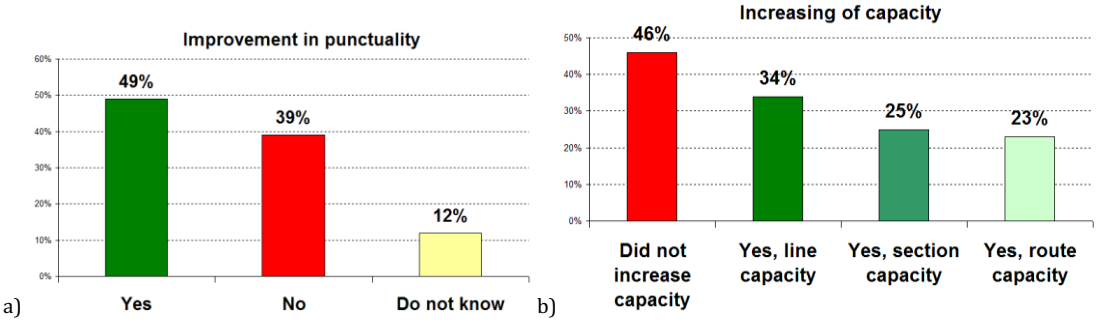


Figure 3: Distributions of the answers to questions: a) „ *Do you think that introduction of the system improved the punctuality of trains?"*; b) *„Do you think the introduction of the system increased the capacity?"* [Her13, Kar12b].

As many as 46% of the respondents see no relation between the speed at which the information is provided and the efficiency of traffic management, leading to the high traffic smoothness and to maintaining the current capacity reserves (Fig 3b). This result may be due to two factors. The first one (and the most important one) is the lack of data in the system, such as an accurate and current information about the position of the train and real times of trains passing the individual stations. The second factor is the potential inability of the users to utilise the information provided and the lack of trust for the data. Furthermore respondents do see the opportunity of using the information provided by the system to increase the capacity of the elements of railway network (of a route – 23% of answers, of a section – 25% of answers and of a line – 34% of the answers). The total percentage in not 100% as this question was of a multiple choice nature [Her13, Kar12b].

The advantages of the system in increasing speed of access to the information were appreciated by as many as 90% of the respondents (the answers included 77% of 'Largely Increased' and 13% of 'Somewhat increased' answers – Fig 4a). Unfortunately, the remaining 10% of the respondents do not see any advantages of an efficient access to actual information. Until now a train dispatcher could obtain all the information on a specific train (route, carrier, scheduled departure, scheduled passing time etc.) by phone to a relevant line controller. [Her13, Kar12b]. Train dispatchers on their shifts make a number of traffic-related decisions on train receiving, dispatching and managing the traffic on adjacent routes. The respondents have rated the comfort of working with the system high and of decision-making (82% of respondents). The remaining 18% of respondents were of an opposite opinion (Fig 4b) [Her13, Kar12b].

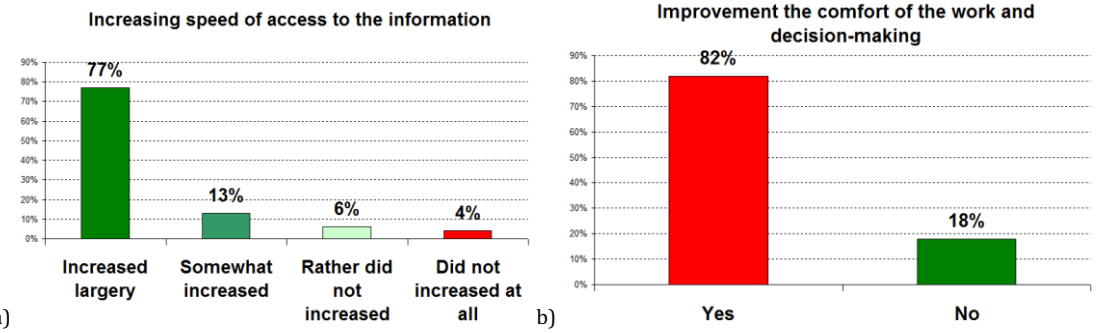


Figure 4: Distributions of the answers to questions: a) „*Did the introduction of the system increase the speed of access to the needed information about a train?"*; b) „*Did the system improve the comfort of your work and of the decision-making?"* [Her13, Kar12b].

The responses at the question "*Is the level of access to various system tabs and the possibilities of their edition satisfactory at your work position?*" indicate that the opinions on that point are divided. Only 36% of respondents were satisfied while 54% assessed the access to the data as unsatisfactory (Fig 5a). This may be due to several reasons. The first reason may be the workplace structure of the respondents. The participants of the survey were the personnel of large traffic control stations for which more access to the system would mean more efficient job, hence the 'unsatisfactory' responses were most frequent in that group. For the system users from small stations, just the basic view of the data is satisfactory. In addition some of them could fall into the group of users who responded 'do not know' (10%). [Her13, Kar12b].

Moreover, the definitive majority of respondents consider an extended access to the system to be helpful. The analysis of the open answers has shown that [Her13, Kar12b]:

- in 80% cases 'the possibility of inputting a real train passing time' was indicated,
- almost 67% of respondents indicated the drawback to be *'lack of possibilities of inputting an analysis of a train dispatched'*. The analysis of a dispatched train requires the collaboration with the line controller and proceeds as follows [Her13, Kar12b]:
    - the train operator calls into the train dispatcher, providing the filled in *The list of wagons in the train.* This list comprises the detailed data on the train,
    - the train dispatcher checks in the SWDR system whether that train has been planned and if its parameters (the serial number and the type of the locomotive, gross weight, train length) are aligned with the plan,
    - the train dispatcher copies the train data into the *Train Analyses Journal,*
    - the train dispatcher contacts the line controller, providing him with the data and informing him about potential delays and their causes,
    - the line controller inputs the information provided by the train dispatcher into the system,

- it is therefore clear that this set of time consuming activities could be largely modified and simplified and partially eliminated.

Before the system was installed and configured on the computers of the train dispatchers, only a short version of user manual was provided to the work posts (either in paper or electronic format). The information in the manual was limited to the general description of the functions and the procedure of logging into the system. In 25% cases the information and knowledge on system use was being obtained from the co-workers, during the breaks in train traffic operations (Fig 5b).

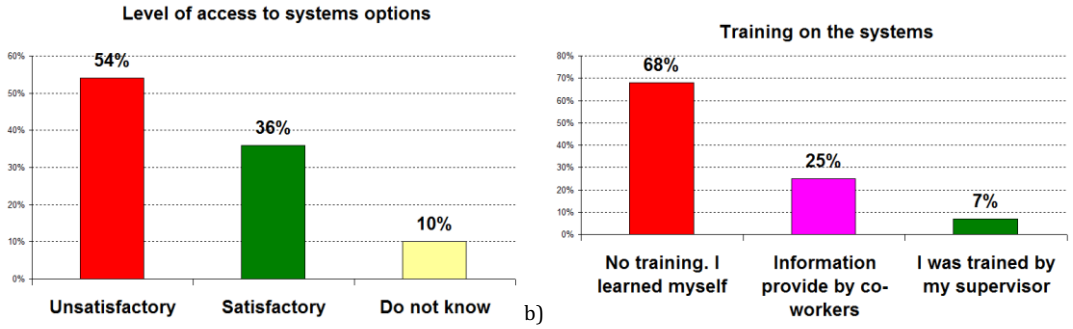


Figure 5: Distributions of the answers to questions: a) *„Is the level of access to various system tabs and the possibilities of their edition satisfactory at your work position?"*; b) *„Have you done any training concerning the operation of these systems?"* [Her13, Kar12b].

The remaining percentage of the respondents (68%) declared they have learned to use the system by themselves. Such a way of deploying a new system certainly does not help the knowledge of the way it needs to be operated and does not provide knowledge on the system functions. And therefore in the next question the need to be trained on the use of the system was raised by 70% of the respondents. This may be a result of a purely technical approach to the application - users know where and when to click. However not always the user of the system is fully aware of the possibilities of using the data provided in the subsequent traffic operations. Thus, a training cycle should not be limited to passing basic information on the use of the application but also should include the studies of concrete examples of practical application of the information available in the system [Her13, Kar12b].

4 Conclusions

The analysis and evaluation of the functionality of the railway ITS systems in relation to the duties of train dispatchers and railway traffic controllers permits drawing the following conclusions [Her13, Kar12b]:

- from the functionality point of view the systems perform well and the majority of functions available corresponds to the needs of the users,
- some modifications are recommended to allow enhanced access to the system for individual users,

- data acquisition from the operating environment should be automated to a higher extent, so that the unnecessary intermediating links are eliminated,
- introduction of changes in the mutual communication of traffic controllers and train dispatchers will improve the ergonomic parameters of their jobs,
- systems integration into one application will make the operation of the system and the information processing more simple; it will also enhance the connectivity with the external transponders; this conclusion is based on results of open–ended questions [Her11a, Her11b] not included in this paper and other own research and analysis,
- effective utilisation of the systems supporting the work of train dispatchers and railway traffic controllers requires that the software implementation is supported by a system of professional trainings,
- the implementation of subsequent system versions should be preceded by direct consultations with the users, who possess the best knowledge about the current problems and the requirements of the system users in daily operations.

The proposed directions of changes within the discussed systems and in their user interface layer may be viewed in two dimensions: a technological one and a procedural one. In the technological change category, the important elements are the need to provide the railway tracks with the track equipment transmitting data signals to the server or alternatively equipping all the rail engines with GPRS (GSM-R) transmitters. These changes are also related to the European Railway Traffic Management System project (ERTMS), which has a GSM-R, telecommunication network which will provide digital radio communication. The category of procedural changes includes the enhanced access to the system for the train dispatchers. Some of the postulated changes have already been included in the system version 2.2 of February 2011 and some other changes will appear in the newest version. These changes allow direct input of the data which previously were being passed by the dispatcher to the traffic controller over the phone (Fig 1). It is also important to remember about the additional equipment and the modification of the workplace (some of these were previously not categorized – both financially and technically – as computerised positions). All of these measures should allow breaking the staff resistance towards the new technology and change the perception of the change meaning the additional workload. [Her13, Kar12b]

References

[FHW07] U. S. Department of Transportation: *Systems Engineering for Intelligent Transportation Systems*. FHWA-HOP-07-069, Jan. 2007.

[FHW06] U. S. Department of Transportation: *Regional ITS Architecture Guidance*. FHWA-HOP-06-112, July 2006.

[Her11a] G. Herzyk and G. Karon: "Funkcjonalność systemów wspomagania pracy dyspozytora liniowego i dyżurnego ruchu w świetle wyników badań ankietowych

z 2010 r". In: *TTS Technika Transportu Szynowego* 11.1-2 (Jan./Feb. 2011), pp. 59–65.

[Her11b] G. HERZYK and G. KARON. "Propozycje zmian w funkcjonalności systemów wspomagania pracy dyspozytora liniowego i dyżurnego ruchu na podstawie wyników badań ankietowych z 2010 r". In: *TTS Technika Transportu Szynowego* 11.3 (Mar. 2011), pp. 54–61.

[Her13] G. HERZYK, G. KARON, and J. MIKULSKI: "The functionality problems of the ITS systems supporting rail transportation – survey results 2010-12". In: *Archives of Transport Systems Telematics* 6.2 (May 2013), pp. 183–198.

[Kar11] G. KARON and J. MIKULSKI: "Transportation Systems Modelling as Planning, Organisation and Management for Solutions Created with ITS". In: *Modern Transport Telematics, CCIS 239.* Ed. by J. MIKULSKI. Berlin, Heidelberg: Springer, 2011, pp. 277–290.

[Kar12a] G. KARON, J. PAWLICKI, and G. HERZYK: "The Characteristics of the Systems Supporting Railway Transportation Process". In: *Contemporary Transportation Systems. Selected Theoretical and Practical Problems.* Ed by R. JANECKI and S. KRAWIEC. Gliwice, Poland: Wyd. Politechniki Śląskiej, 2012, pp. 209–218. ISBN: 978-83-7335-944-4.

[Kar12b] G. KARON, J. PAWLICKI, and G. HERZYK. "Functional Evaluation of the Information Systems Supporting Train Dispatchers and Raiway Traffic Controllers – Based on 2010". In: *Contemporary Transportation Systems. Selected Theoretical and Practical Problems.* Ed by R. JANECKI and S. KRAWIEC. Gliwice, Poland: Wyd. Politechniki Śląskiej, 2012, pp. 219–232. ISBN: 978-83-7335-944-4.

[Kar12c] G. KARON and J. MIKULSKI: "Problems of ITS Architecture Development and ITS architecture implementation in Upper-Silesian Conurbation in Poland". In: *Telematics in the Transport Environment, CCIS 329.* Ed. by J. MIKULSKI. Berlin, Heidelberg: Springer, 2012, pp. 183–198.

[Kar12d] G. KARON and J. MIKULSKI. "Modelling of Transportation Systems in Agglomeration for ITS Solutions". In: *CESCIT 2012 1st IFAC Conference on Embedded Systems, Computational Intelligence and Telematics in Control.* Ed. by K. SCHILLING and F. LEUTERT. Wurzburg, Germany: Universitat Wurzburg, 2012, pp. 74–79.

Corresponding author: Grzegorz Karon, Silesian University of Technology, Faculty of Transport, Krasinskiego 8 str. 40-019 Katowice, Poland, phone: +00 32 603 4 159, e-mail: grzegorz.karon@polsl.pl

RTSE, a Multi-Component Closed-Loop Control Framework for Railway Networks

Raimond Wuest[1], Albert Steiner[1], Jonas Looser[1], Bernhard Seybold[1], Marco Laumanns[2], Juliane Dunkel[2], Daniel Huerlimann[3], Samuel Roos[4]

[1] Zurich University of Applied Sciences

[2] IBM Research - Zurich

[3] OpenTrack Railway Technologies Ltd.

[4] Emch + Berger AG

Abstract

Optimal operation of rail transport systems has become an increasingly challenging task over the last decades. To allow for a better understanding of the system dynamics in different operational states (including disruptions) and in order to evaluate and to improve control strategies, a multi-component simulation framework, representing a closed-loop operation environment for railway networks, is being developed. This framework is based on a time controlled and partially automated operational concept. Time control requires all operational processes to be continuously monitored with respect to the production schedule. Deviations exceeding some pre-determined tolerance thresholds will result in a re-adjustment of the production plan in real-time. A dedicated (re-)scheduling algorithm is implemented to achieve this goal.

Involved parties (agents) are explicitly taken into account. For instance, train drivers might be technically enabled to follow new operational targets like re-adjusted train speeds while approaching conflict points. The framework, called Rail Transport Service Environment (RTSE), consists of three main modules: (i) a traffic simulation environment, (ii) a system state monitoring module, and (iii) the scheduling module. The modules are interconnected through standard communication interfaces so that each module can be exchanged easily depending on the user environment. Railway traffic simulations are carried out using the dedicated railway simulation tool OpenTrack. The simulated traffic situations are interpreted by an automated monitoring module including a threshold detection mechanism, which compares actual and planned process states and induces rescheduling actions executed by the (re-)scheduling algorithm, if required. Rescheduling actions take eventually reduced availability of resources into account.

Keywords: Rail traffic – Rail traffic simulation– Rail traffic state monitoring – Real-time dispatching – Network performance – Service intention

1 Introduction

Modern railway operational processes abandoning accumulated legacy cases, explicit description of the service to be delivered to the customer, and increased computational and communication performance allow to make rail transportation service more predictable as well as to orient operations to customer benefit even in case of incidents. Operation according to plan even in case of delay or disruption allows for a better use of contested capacity. With our proposed framework we aim to demonstrate and evaluate this opportunity.

1.1 Motivation

The identification of increasing capacity problems that ask for a redesign of railway operations is the main motivation of the project participants for developing the proposed framework. In order to derive the elements of the proposed approach, we will first have a closer look at the major problems.

Challenges in service delivery

The Swiss railway network and the services delivered through it can be characterized as a multiple hub-and-spoke network with integrated clock-face timetable and it is well known to be strongly interconnected. This means that there are lots of point to point services that require one or more train transitions. In these cases individual delays often have impacts on a large part of the network. This is the case if train dispatchers decide to hold back the connecting trains affected or if they decide not to keep communicated connections.

On the other hand, the operations staff as well as the customers are facing typical problems that result from the lack of timely information on the system as illustrated in Figure 1.

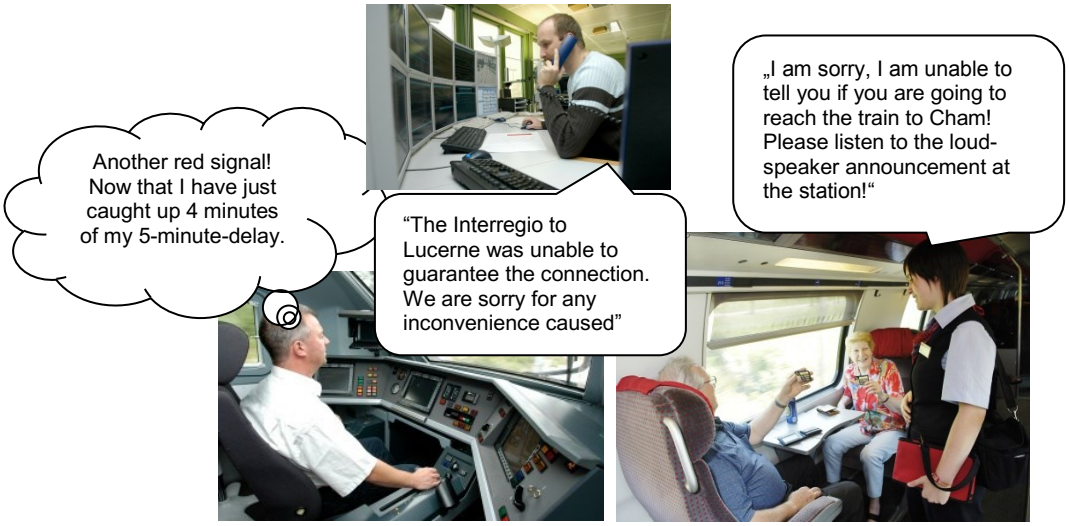


Figure 1: Operating staff and customers confronted with decreasing service reliability, each with different responsibilities within the service process chain.

Following the above explanations, the major service delivery challenges can be summarized as follows:

- a tight regular timetable due to an increasing gap between peak hour and off peak hour demand.
- decreasing service reliability due to operational volatility and technical disturbances.
- limited usability of public transport due to communication problems.
- significant total passenger delays due to local dispatching decisions.

If we analyse specific cases and try to figure out what happened before the occurrences of major network delays, we observe similar patterns. We see that in most cases operational disturbances lead to blocked resource assignments and hence unusable production plans. As the production plan is the technical basis for operating the timetable, it is the main task of operational control to reassign resources such that the normal timetable and the planned services will be restored as soon as possible. This is a complex task that requires, even in the case of small delays, the consideration of numerous operational and technical constraints.

2 Solution approach

2.1 Service Intention

Usually timetable development is an extensive, iterative planning process starting years ahead of the beginning of the actual timetable period. The main objective is to find train runs with departure, arrival and dwell times that simultaneously meet functional requirements (Service Intention, SI) connecting each origin and destination in the network with required travel times, frequencies and service levels as well as technical requirements. Technical constraints concern the utilization of resources, such as track topology, rolling stock and operations staff, which partly still have to be implemented for the planned time horizon. The central part of the proposed approach is the assumption that any timetable is one out of several possible technical realizations of an underlying service intention. Because the SI can be regarded as a functional requirement for the scheduling task, it is defined in terms of frequencies between origins and destinations, travel times and service levels rather than of exact departure and arrival times, rolling stock utilization or even trains numbers. Although these attributes are information typically provided by a public transport timetable, it is nothing else than the result of an assignment of processes and resources to these functional requirements, while considering numerous spatio-temporal constraints. As a consequence, operational irregularities or disruptions, which, for a certain period of time prohibit further operation of the timetable planned, require a new assignment of available resources to the functional requirements. To formalize this approach, we define the periodic service intention similar to [Cai09] as

$$SI = (Z, C, D, \bar{\rho}), \tag{1}$$

where Z denotes the set of all train runs z (with $z \in Z$), C is the set of all connections, D denotes the set of all technical and operational dependencies and $\overline{\rho}$ is a given time period (for details see [Cai09]).

2.2 Production plan

Taking the SI defined in (1) as well as available resources (e.g. rolling stock, signal and track topology) as input, a dedicated (re-)scheduling procedure - initially or when triggered by operational disturbances - generates a production plan, which contains all necessary information for operating the train service. Its formal representation is a list of assignments of event times $d_i(k)$ to triples (E_i, L_i, S_i), where E_i represents the event type (arrival, departure, passage), L_i the train line, S_i a location (station, signal, junction) and k the k^{th} occurrence of a periodic event i within a given time period T. In the (re-)scheduler optimization model the SI-requirements enter as constraints. However, as we cannot guarantee that each rescheduling instance will be feasible under the requirements, they are used as a soft constraint, which means we allow constraint violation for a certain penalty. These constraint violations are then penalized in the objective function and the constraint violation penalty is minimized. This general approach includes, e.g., the minimization of knock-on delays as a special case, as a delay is a particular case of a constraint violation. If, in real time, the dispatcher is not satisfied with the resulting solution, she can manually relax some SI-requirement, which will automatically give more flexibility to satisfy other, more important SI-requirements. In this iteration process the (re-)scheduler creates production plans for several variants of SI that are provided by the dispatcher (see process loop in the Management Layer of Figure 3). For each train in the system the production plan contains optimized

- departure times and speed instructions
- route allocations, route reservation and release times
- connections, platform assignments etc. (see Figure 5)

2.3 Process model

Technical and operational dependencies mainly result from allocations of resources (for instance track segments) to operation processes (e.g. departure of train xy). In our framework, planned and logged time extensions of operation processes as well as state transitions of affected resources are modeled as 'Timed Event Graph' (TEG). For instance, safety blocks are treated as resources which are assigned to train runs by the (re-)scheduler. A typical snapshot of this process model is shown in Figure 2 as a timed petri net representation of the departure process of a train run. The benefit is twofold. On one side, we have the perfect structure to monitor time stamps of planned events with those recorded by the system (mainly the safety system) or human actors. As this method is robust against the exact order of message occurrence we can also find out in this way if a planned event is overdue.

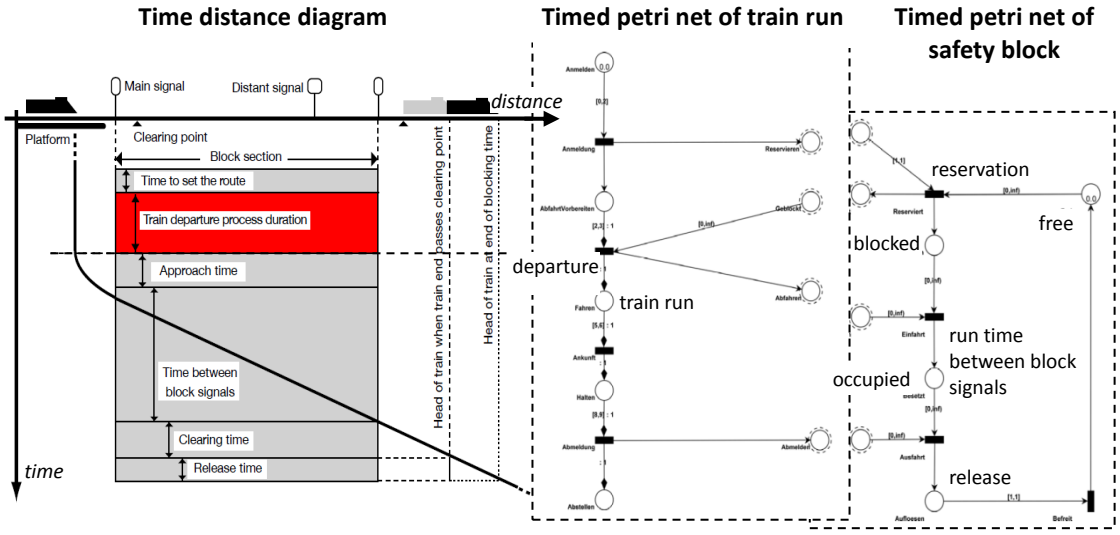


Figure 2: Integrated process model: example of departure process and its representation as timed event graph.

On the other side, we are able to forecast the system behavior by evaluating the process dependencies in the model - assuming that all process will execute as planned due to the information in the output of the rescheduling procedure. We do this by implementing the delay propagation procedure according to [Gov10] extending the method to the resources involved. This framework also enables us to analyze timetable stability, e.g. using max-plus algebra according to [Gov07].

3 System design

To address the problems outlined in chapter 1 and to incorporate the fundamental conceptual elements described in detail in chapter 2, we developed the general framework shown in Figure 3. The system consists of three layers, each with dedicated functionalities.

The *Management Layer* is at the top. One of its components is the dispatcher who represents the highest decision level in the *Management Layer*. The dispatcher's task is to manage the functional requirements, i.e. the SI. The SI is entered into the (re-) scheduler component of the *Management Layer* (interface number 1) as a basis for the calculation of the normal production plan. In case of an operational rescheduling requirement that cannot satisfy original SI-requirements, the dispatcher has to relax the SI such that the (re-) scheduler can find a feasible solution for a new temporary production plan (interface 4). The production plan (see 2.2) is displayed either as a list (see Figure 4 a) or as a timed petri net (see Figure 4 b).

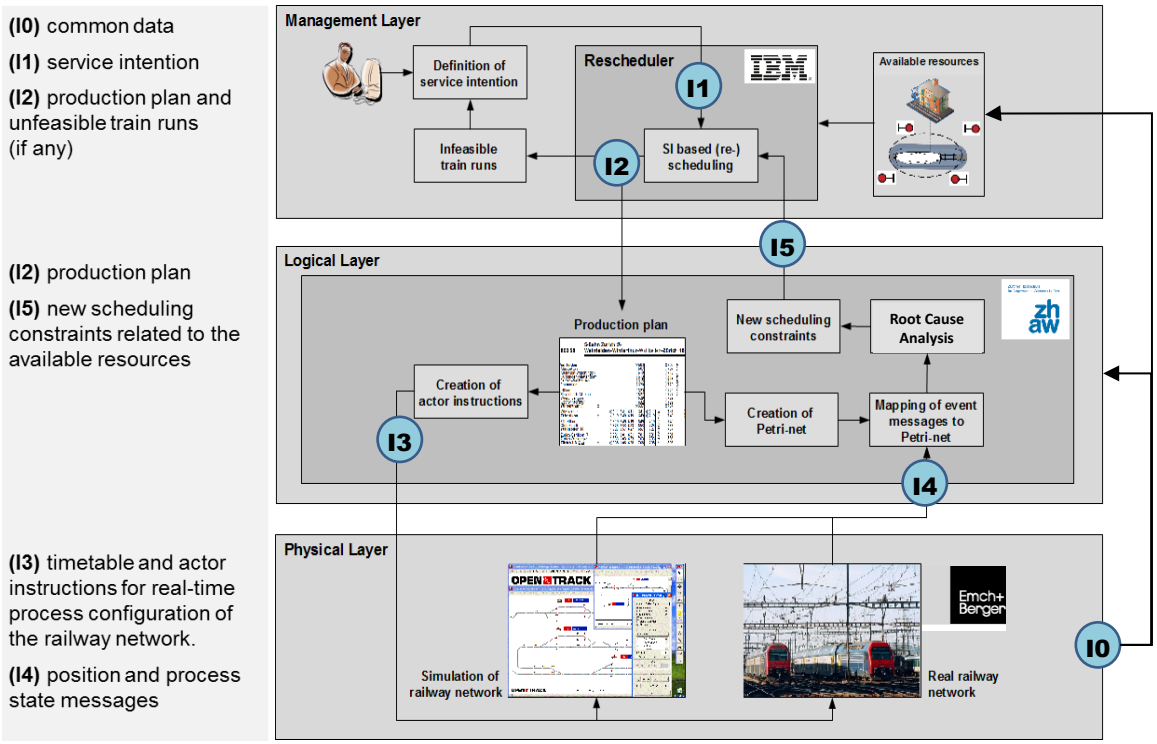


Figure 3: The functional components of the system, the feedback loop and the corresponding interfaces (interface numbers are shown in blue circles): (I0) configuration data, (I1) service intention, (I2) production plan and unfeasible train runs (if any), (I3) timetable and actor instructions for real-time process configuration, (I4) position and process state messages, (I5) new scheduling constraints related to the resources available.

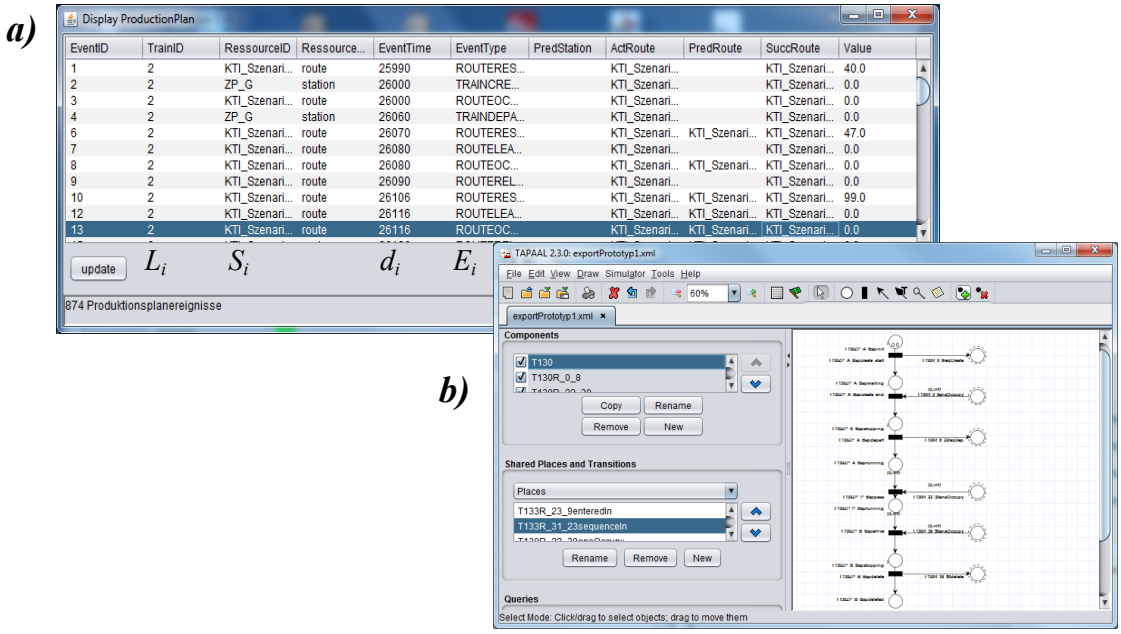


Figure 4: production plan representation a) list, b) timed petri net.

After actor specific formatting, the production plan is used by the *Logical Layer* for (re-) configuration of the railway network. During run time, position and process state messages generated in the railway network (interface 4) are evaluated permanently by the operation control part in order to detect threshold exceedings and new constraints and to initiate rescheduling (interface 5) in the *Management Layer*. The message interfaces of the *Logical Layer* (interfaces number 2, 3, 4 and 5) are easy to standardize. This ensures that conclusions that can be drawn from the simulation environment are transferable to real world conditions. On the other hand, different rescheduling modules (*Management Layer*) can be used without changing anything in the Logical Layer.

Components of the *Logical Layer* execute (simulate) physical processes and generate messages with time stamps that are supposed to correspond to those of the production plan (interface 3, see Figure 5).

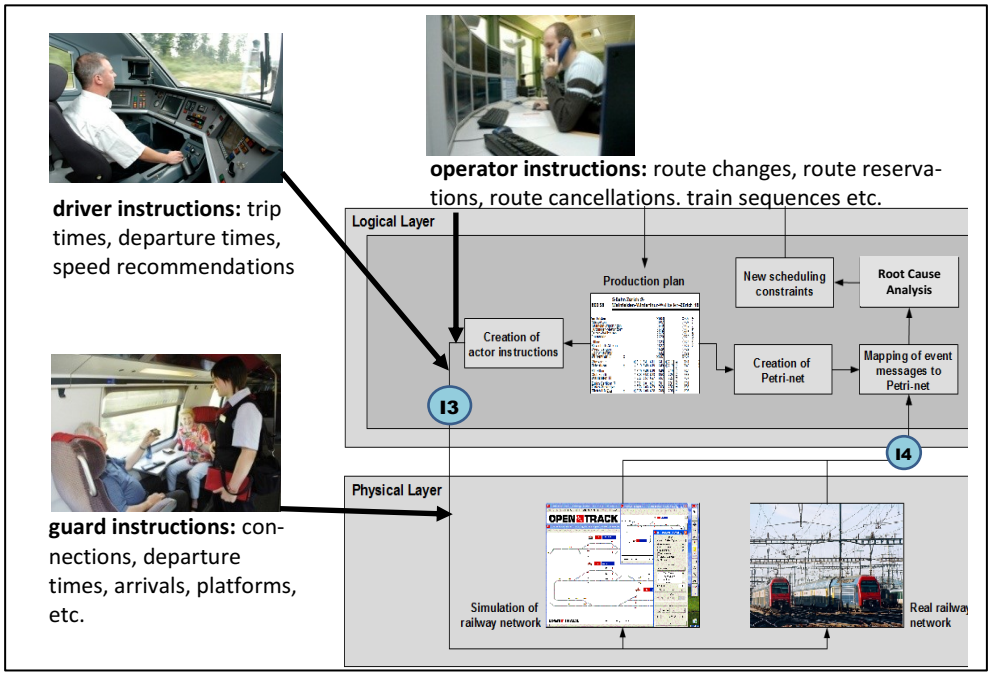


Figure 5: Actor instructions to train drivers, operators (resp. traffic management system) and train guards affected by the (re-)scheduling action (interface I3).

The *Physical Layer* consists of the operational environment, represented either by the simulation environment (in our case the railway simulation tool OpenTrack [OpT12]) or by the real-world system. It carries out all transport operation, safety, customer and disruption processes and contains detailed information about resources (interface 0).

4 System behaviour

The system behaviour is strongly influenced by the precision, with which the dynamics of train runs can be approximated by the (re-)scheduling algorithm of the Management Layer. But as the behaviour of the simulated train runs is only controlled via data configuration and message instructions (train departure, train speed etc.) it is essential to generate the right speed instructions at the right time.

An example of an intercity train run in the given test scenario (see Figure 6) is illustrated in Figure 7. The train dynamics calculated in OpenTrack are considered to be realistic. The shown data result from a simulation with a train run that was completely controlled by instructions of the production plan (see Figure 5). The plotted event time deviations result from system delay resp. earliness due model discrepancies between (re-)scheduler and OpenTrack (route reserve commands are executed immediately if safety blocks are free). It shows that the precision of the dynamic behavior of our framework is reasonable in normal (undisturbed) situations. Deviations above threshold are corrected instantaneously by a new production plan which again results in undisturbed conditions for a certain period (closed control loop). In the RTSE rescheduling module we use the model from Caimi et al 2012 and solve it with the MIP solver IBM ILOG CPLEX Version 12.5.1 (IBM, ILOG, and CPLEX are trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies).

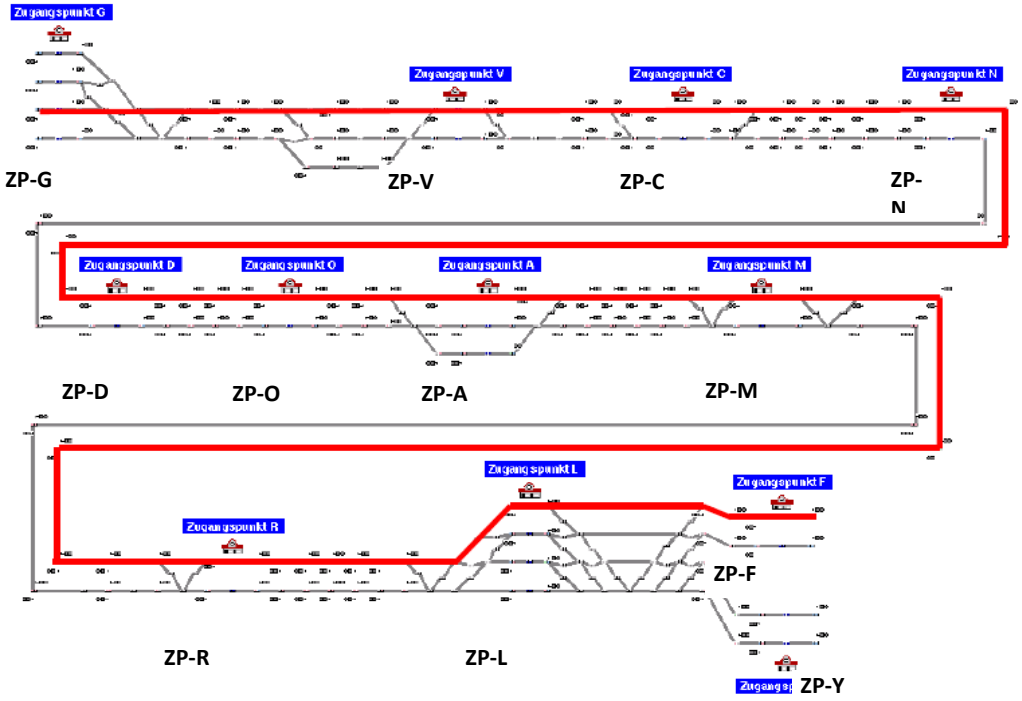


Figure 6: Topology of test scenario with 12 stations. Path of example train in red.

5 Conclusion and outlook

The behaviour of the RTSE framework, which we present in this paper is of sufficient accuracy to simulate train operations in a closed loop. Our next steps will be to implement the entire system setup in an online architecture and to use this to investigate the dynamic closed loop behaviour of a railway network in real time. Thus we will establish a system environment that allows us to benchmark dispatching decisions in terms of good choices for relaxed SI's under conditions of disturbed train operations.

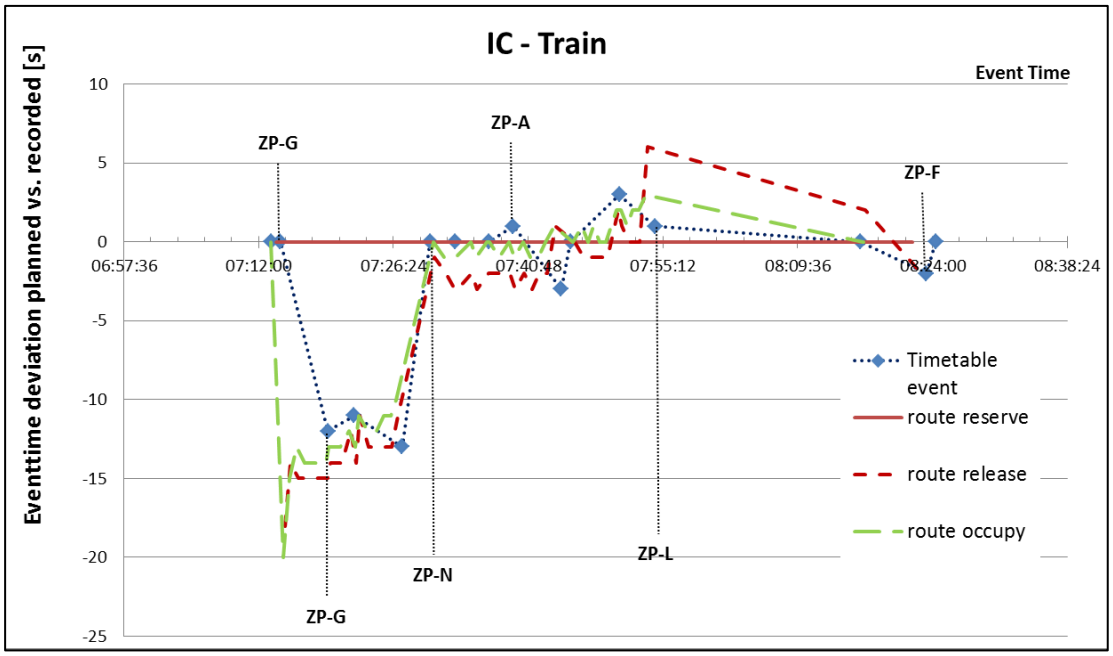


Figure 7: Deviation of planned vs. recorded event times of intercity train run.

References

[Cai09] G. C. CAIMI: "Algorithmic decision support for train scheduling in a large and highly utilised railway network". PhD Thesis. Zurich: Institute for Operations Research, ETH Zurich, 2009, pp. 52ff.

[Cai12] G. C. CAIMI, M. FUCHSBERGER, M. LAUMANNS, and M. LUETHI: "A model predictive control approach for discrete-time rescheduling in complex central railway station areas". In: *Computers & Operations Research* 39.11 (2012), pp. 2578–2593.

[Gov07] R. M. P. GOVERDE: "Railway timetable stability analysis using max-plus system theory". In: *Transportation Research Part B: Methodological* 41.2 (2007), pp. 179–201.

[Gov10] R. M. P. GOVERDE: "A delay propagation algorithm for large-scale railway traffic networks". *Transportation Research Part C: Emerging Technologies* 18.3 (2010), pp. 269–287.

[OpT12] OPENTRACK RAILWAY TECHNOLOGY GMBH: *OpenTrack.* (Last Access: 29 April 2012). URL: `http://www.opentrack.ch`.

Corresponding author: Raimond Wuest, Zurich University of Applied Sciences, Institute for Data Analysis and Process Design, CH-8401 Winterthur, Switzerland, phone: +41 58 934 6581, e-mail: raimond.wuest@zhaw.ch

A Passenger Knock-On Delay Model for Timetable Optimisation

Peter Sels[1, 2, 3], Thijs Dewilde[1], Dirk Cattrysse[1], Pieter Vansteenwegen[1]
[1]KU Leuven, University of Leuven
[2]Logically Yours BVBA, Antwerp
[3]Infrabel, Department of Rail Access, Brussels

Abstract

In the process of timetable creation, sufficient time should be scheduled between any pair of trains using a common infrastructure section in order to avoid that a delay on the first train will cause a delay on the second train too. However, when this time buffer becomes very high, the positive incremental buffering effect diminishes and other negative effects may appear, like reduced timetable efficiency or lower than optimal remaining time between the other trains on the same infrastructure resource. This means there is a trade-off to make. We make this trade-off by analytically deriving the knock-on delays as passengers experience it in practice and by including these delays in our goal function: total expected passenger journey time in practice.

We use this goal function in our Mixed Integer Linear Programming (MILP) model to optimise from scratch, the timetable of all 203 hourly passenger trains in Belgium. We then also compare our resulting timetable with the original schedule, and conclude that both the knock-on component as well as the total expected passenger time are reduced.

Keywords: Knock-On Delay Model, Expected Passenger Time, Integer Linear Programming, Goal Function

1 Introduction

A railway timetable can be aptly represented by a graph. Graph vertices are train arrival and departure times. The graph's edges are either primary edges representing intra-train actions: ride and dwell, or secondary edges, representing inter-train actions: transfer or turn-around. Other secondary edges represent a required time difference: headway requirements [Kro09]. For all edges, primary or secondary, a minimum time is required and we also add a non-negative supplement. Note that we use the term *supplement* also in the meaning of *buffer* between two trains on a common infrastructure resource. The purpose of the supplements can

be twofold. First they are sometimes needed as slack between two already planned timetable times. Indeed, imagining that one would plan the primary edges first, some slack would result for the inter-train transfer, turn-around and headway edges. Secondly a larger than slack-only supplement could be needed to make a timetable robust against delay. However, supplements may also not become too large, resulting in trains riding too slow or idling too much and as such resulting in an inefficient planning. So, obviously there is a trade-off, per supplement, between robustness and efficiency. Additionally, when edges are part of a common graph cycle, the sum of minimum process times and supplements over all edges of the cycle have to sum up to a multiple of the timetable period [Gov10]. This means choices of supplements of these edges are related and one also has to be able to properly weigh the costs and benefits of the supplement choices on different edges. We consider one train more important than another when it has more passengers present on it. We could introduce artificial train class priorities, but prefer to directly weigh importance with passenger numbers instead. In [Sel11] we derived passenger numbers on all trains at all locations, starting from ticket sales data. With this information, we can formulate the *total expected passenger time in practice* [Dew11; Sel13b] as a function of the timetable. More specifically, it is a function of 3 parameters sets: (1) the action minima, (2) the assumed primary delays and (3) the planned supplements. Secondary delays also increase this expected passenger time, but are itself a function of the three mentioned parameter sets. The resulting total function is to be minimised to generate an optimal timetable for passengers. The minima are fixed, so in each timetable it will generate the same amount of expected time. The supplements are the decision variables of the timetable, so given the delay assumptions, their values determine any quality criterium of the timetable as expected passenger time, robustness and efficiency.

The total expected passenger time has been analytically derived as a function of minima and supplements in [Sel13b] for departing, through, transfer and arriving passengers. In this paper we add the derivation of the knock-on delay as a function of the minimum and supplement present on a headway edge. Indeed both a headway minimum time as well as a knock-on delay should be modelled whenever two trains on a common resource occur. So a *hard* headway constraint and a *soft* knock-on cost as a term in the goal function are always modelled on the same edge.

Section 2 lays out an analytical derivation of the knock-on delay function. Section 3 presents the results obtained when using these knock-on delay functions as terms in the goal function for a system of all 203 trains currently departing between 7 and 8 am in the cyclic Belgian timetable. Section 4 draws conclusions and hints at some further work.

2 Knock-On Delay Derivation

When train i is riding or dwelling on a track and it gets delayed, it can delay train j which follows it on the same track. We will derive a cost function that gives us the expected delay for all passengers on the second train as a function of the planned time in between the two trains and the expected delays on these trains.

We define the number of passengers on train i as f_i and on train j as f_j. As [Sel13a] explains, for trains riding in the same direction on a common track, headway edges exist between both the vertices representing the beginning of the trains' ride actions in both directions, cyclically and also between the endings of the trains' ride actions again in both directions, cyclically. For trains riding in opposite directions on a common track, a headway edge exists between the end of the first train's ride action and the beginning of the other train's ride action and vice versa, cyclically. In the sequel, when we mention a knock-on edge between train i and j, we more specifically mean the knock-on edge between two vertices v_i and v_j, where these vertices can be a begin or end vertex of a ride edge.

We can suppose the vertices v_i and v_j, which represent event times, to experience primary delays according to (commonly used [Han08]) negative exponential distributions

$$p_i(x) = a_i e^{-a_i x}, p_j(y) = a_j e^{-a_j y}, \tag{1}$$

where x and y are the primary delays of time points v_i and v_j and $p_i(x)$ and $p_j(y)$ their respective probabilities. The *expected* delays of these distributions are calculated to be

$$\overline{c_i} = \int_0^\infty x a_i e^{-a_i x} = \frac{1}{a_i}, \overline{c_j} = \int_0^\infty y a e^{-a_j y} = \frac{1}{a_j}. \tag{2}$$

Say that, on top of the mandatory heading time h between trains i and j, which has to be respected at any time, there is a *planned* supplement time $s_{i,j}$ and similarly a *planned* supplement $s_{j,i}$ between trains j and i. Then, the probability that due to combined delays of trains i and j one train will delay the other is calculated by adding all cases where the delay difference of both trains exceeds the supplement between them, weighting these cases with the probability that they occur. This is done by integrating over a triangle area where the delay difference $x - y \geq s_{i,j}$ so $x \geq y + s_{i,j}$ and over another where $y \geq x + s_{j,i}$ as in

$$\begin{array}{rcl} p_{x \geq y + s_{i,j}}(a_i, a_j, s_{i,j}) & = & \int_0^\infty \int_{y+s_{i,j}}^\infty a_i e^{-a_i x} \cdot a_j e^{-a_j y} dx dy = \frac{a_j e^{-a_i s_{i,j}}}{a_i + a_j}, \\ p_{y \geq x + s_{i,j}}(a_i, a_j, s_{j,i}) & = & \int_0^\infty \int_{x+s_{j,i}}^\infty a_i e^{-a_i x} \cdot a_j e^{-a_j y} dy dx = \frac{a_i e^{-a_j s_{j,i}}}{a_i + a_j}. \end{array} \tag{3}$$

In the area where $x < y + s_{i,j}$ and $y < x + s_{j,i}$, $s_{i,j}$ respectively $s_{j,i}$ are large enough to absorb primary delays and avoid knock-on delays. The total *expected* knock-on delay of train i on train j is calculated by multiplying, for each case where a knock-on delay occurs, its probability, with the knock-on delay amount occurring and then integrating these products over the same triangular integration areas as before. Via partial integration, one can prove

$$\begin{array}{rcl} tKO_{i,j}(a_i, a_j, s_{i,j}) & = & \int_0^\infty \int_{y+s_{i,j}}^\infty a_i e^{-a_i x} \cdot a_j e^{-a_j y} (x - y - s_{i,j}) dx dy = \frac{a_j e^{-a_i s_{i,j}}}{a_i (a_i + a_j)}, \\ tKO_{j,i}(a_i, a_j, s_{j,i}) & = & \int_0^\infty \int_{x+s_{j,i}}^\infty a_i e^{-a_i x} \cdot a_j e^{-a_j y} (y - x - s_{j,i}) dy dx = \frac{a_i e^{-a_j s_{j,i}}}{a_j (a_i + a_j)}. \end{array} \tag{4}$$

From equations (4), two properties can be derived. First, the larger the planned separation time $s_{i,j}$ between the trains, the lower $tKO_{i,j}$, so the lower the expected knock-on delay on train j. Second, the lower the expected primary delay $\overline{c_i} = 1/a_i$ on train i, the higher a_i,

the lower $tKO_{i,j}$, so the lower the expected knock-on delay on train j. These tendencies are indeed what we expect in practice as well. Since we are interested in the knock-on delays as *passengers* experience them in practice, we multiply the train knock-on delay with the number of passengers on the knocked-on train and get

$$\begin{aligned} pKO_{i,j}(a_i, a_j, s_{i,j}) &= f_j \cdot tKO_{i,j} = f_j \cdot \tfrac{a_j e^{-a_i s_{i,j}}}{a_i(a_i+a_j)}, \\ pKO_{j,i}(a_i, a_j, s_{j,i}) &= f_i \cdot tKO_{j,i} = f_i \cdot \tfrac{a_i e^{-a_j s_{j,i}}}{a_j(a_i+a_j)}. \end{aligned} \tag{5}$$

If only two trains i and j are to be planned on a common resource, in a one hour period, what are the ideal supplement times $s_{i,j}, s_{j,i}$ to be planned in between them? This will depend on their passenger numbers f_i, f_j and their expected delays a_i and a_j. First, note that there is a relation to respect between $s_{i,j}$ and $s_{j,i}$. Indeed, the constraint for the cycle formed by the two headway edges between trains i and j is

$$h + s_{i,j} + h + s_{j,i} = T \text{ or equivalently } s_{j,i} = T - 2h - s_{i,j}. \tag{6}$$

After substitution of $T - 2h - s_{i,j}$ for $s_{j,i}$ in $pKO_{j,i}$, $pKO_{j,i}$ is clearly a function of $s_{i,j}$. Since $pKO_{i,j}$ and $pKO_{j,i}$ are both convex functions of $s_{i,j}$, their sum is a convex function of $s_{i,j}$ as well. This means the optimal spreading of two trains per time period T can be calculated by minimising the total expected delay on all passengers of both trains as

$$\begin{aligned} 0 &= \tfrac{d}{ds_{i,j}} \left(pKO_{i,j} + pKO_{j,i} \right) \\ \Leftrightarrow 0 &= \tfrac{d}{ds_{i,j}} \left(f_j \cdot \tfrac{a_j e^{-a_i s_{i,j}}}{a_i(a_i+a_j)} + f_i \cdot \tfrac{a_i e^{-a_j(T-2h-s_{i,j})}}{a_j(a_i+a_j)} \right) \\ \Leftrightarrow 0 &= -f_j \cdot \tfrac{a_j e^{-a_i s_{i,j}}}{a_i+a_j} + f_i \cdot \tfrac{a_i e^{-a_j(T-2h-s_{i,j})}}{a_i+a_j} \\ \Leftrightarrow f_j \cdot a_j e^{-a_i s_{i,j}} &= f_i \cdot a_i e^{-a_j(T-2h-s_{i,j})} \\ \Leftrightarrow ln\left(\tfrac{f_j \cdot a_j}{f_i \cdot a_i}\right) &= -a_j(T-2h-s_{i,j}) + a_i(s_{i,j}) \\ \Leftrightarrow s_{i,j} &= \tfrac{a_j(T-2h) + ln\left(\frac{f_j a_j}{f_i a_i}\right)}{a_i+a_j} \end{aligned} \tag{7}$$

It follows from symmetry that

$$s_{j,i} = \frac{a_i(T-2h) + ln\left(\frac{f_i a_i}{f_j a_j}\right)}{a_i + a_j}. \tag{8}$$

The right hand sides of equations (7) and (8) sum up to $T - 2h$ as equation (6) requires.

As an example, for $T = 60$ minutes and $h = 3$ minutes, a train i with an expected delay of $1/a_i = 3$ minutes and $f_i = 100$ passengers on it and a train j with an expected delay of $1/a_j = 1$ minute and $f_j = 300$ passengers, would be spread according to equations (7) and (8) as $s_{i,j} = \frac{1(60-2\cdot3)+ln(300\cdot1/(100\cdot1/3))}{1/3+1} = 42.15$ minutes and $s_{j,i} = \frac{1/3(60-2\cdot3)+ln(100\cdot1/3/(300\cdot1))}{1/3+1} = 11.85$ minutes and indeed as equation (6) requires $42.15 + 3 + 11.85 + 3 = 60$ minutes.

This kind of balancing of supplements between trains on the same resource will be done by our solver when we add the costs in equation (5) to the goal function. (Note that also choices of supplements on graph edges in common cycles can affect the choice of $s_{i,j}$ and $s_{j,i}$

and vice versa.) We take the approach of generating all knock-on costs between *all* train pairs using the same infrastructure resource, irrespective of their order. This has two reasons. First, unlike the method where we add only knock-on costs between directly subsequent trains, this method works without relying on the yet unknown order of trains. Second, suppose trains i, j and k follow each other in this order on a resource and train i has a large expected primary delay $1/a_i$, train j has a small $1/a_j$ but has very few people f_j on it while train k has a lot of people f_k on it. Then $pKO_{i,j}$ and $pKO_{j,k}$ can be small for low $s_{i,j}$ and low $s_{j,k}$, allowing the three trains, ordered as i, j, k, to be scheduled close together in time, even though $pKO_{i,k}$ will then be large. The fact that cases where $pKO_{i,k} \gg pKO_{i,j} + pKO_{j,k}$ can occur, shows that $pKO_{i,k}$ has to be added to capture all potential knock-on costs.

For N trains using the same resource during every timetable period T cyclically, this method generates $N \cdot (N-1)$ knock-on terms in the goal function. For each resource R, we define the index set I_R as the set of indices of trains that use R. Then, according to equation (5), the total knock-on cost pKO_R for all trains which use resource R is

$$\forall R : pKO_R = \sum_{\substack{i,j \in I_R \\ i \neq j}} f_j \cdot \frac{a_j e^{-a_i s_{i,j}}}{a_i(a_i + a_j)}. \tag{9}$$

For evaluation of the knock-on cost of a given schedule or for non-linear optimisation, equation (9) can be directly used. For a linear solver though, we need to linearise it first. Since each of the terms in (9) is convex in the variable $s_{i,j}$, we can use a standard linearisation trick for convex cost functions. This entails two steps. First, for each of the terms, we define a helper variable $pKO_{R,i,j}$ and impose on them

$$\forall R : \forall_{\substack{i,j \in I_R \\ i \neq j}} : pKO_{R,i,j} \geq f_j \cdot \frac{a_j e^{-a_i s_{i,j}}}{a_i(a_i + a_j)}. \tag{10}$$

All helper variables $KO_{R,i,j}$ are added to the global goal function of expected passenger time. Units match. Since we *minimise* our global goal function, all $KO_{R,i,j}$ are pushed down so that they will be equal to instead of greater than the right hand side of equation (10). Second, the right hand side of (10) is replaced by a number of line segments approximating it. Here, we use 2 segments. So for each $KO_{R,i,j}$ term, we define three points

$$\forall R : \forall_{\substack{i,j \in I_R \\ i \neq j}} : \begin{cases} (s_{i,j,0}, ko_{i,j,0}) = (0, f_j \cdot \frac{a_j}{a_i(a_i+a_j)}) \\ (s_{i,j,1}, ko_{i,j,1}) = (T/15, f_j \cdot \frac{a_j e^{-a_i T/15}}{a_i(a_i+a_j)}) \\ (s_{i,j,2}, ko_{i,j,2}) = (T, f_j \cdot \frac{a_j e^{-a_i T}}{a_i(a_i+a_j)}). \end{cases} \tag{11}$$

The low and high end of the segments are 0 and T so that the whole supplement range is covered. We use $T/15$, or 4 minutes for T equal to one hour, as the abcis of the middle point, because, in our tests, this resulted in the closest approximation to the curve $KO_{R,i,j}$ for most practical cases. Then, with these known values, equation (10) is linearised to

$$\forall R : \forall_{\substack{i,j \in I_R \\ i \neq j}} : \begin{cases} pKO_{R,i,j} & \geq & ko_{i,j,0} + \frac{ko_{i,j,1} - ko_{i,j,0}}{s_{i,j,1} - s_{i,j,0}} \cdot (s_{i,j} - s_{i,j,0}) \\ pKO_{R,i,j} & \geq & ko_{i,j,1} + \frac{ko_{i,j,2} - ko_{i,j,1}}{s_{i,j,2} - s_{i,j,1}} \cdot (s_{i,j} - s_{i,j,1}) \end{cases} \tag{12}$$

We add all $pKO_{R,i,j}$ as variables to our goal function and add the inequalities (12) with the values calculated as in (11) as hard constraints to our MILP model. As such, we have extended our model with a method that accounts for knock-on delays in a way that is properly balanced with the other goal function terms. Note that the obtained estimation of passenger knock-on delay cost can also be used in other than timetable optimisation models. A linear optimisation model maximising capacity consumption with the goal of capacity estimation, as for example [Mus13], could forbid or penalise scenarios with too much knock-on delay.

3 Optimisation Results

For all 203 hourly passenger trains in Belgium, departing between 7 and 8 am in the timetable of June 12th 2013, visiting 1770 open line track sections and calling at all 550 stations, the macroscopic model of constraints as described in [Sel13a] has been set up. (Overtaking is only allowed in stations with 4 or more platform tracks.) The goal function - expected passenger time in practice - as described in [Sel13b] and now extended with the cost terms for knock-on delays, as derived here in section 2, has been constructed. For each ride and dwell action we assumed varying primary delay distributions with an average of $a\%$ of each action's minimum time. a is given in column 1 of table 1. We compare properties of the original and optimised timetable in the next sections.

3.1 Feasibility: A Solution is Always Returned

Since our model has a goal function that properly penalises the choice of big supplements in a soft yet passenger optimal way, there is no reason for us to add a hard constraint that restricts supplements to any arbitrary value lower than $T - \delta$, where δ is the time resolution of the timetabling model. Other research groups (e.g. Delft [Spa13], e.g. Rotterdam [Kro09]) lack a goal function that automatically restricts *all* supplements and so have to enforce lower more arbitrary upper bounds as a hard constraint on their supplements. As a result they sometimes struggle with infeasibility of their model. We believe we have resolved this issue.

3.2 Quality: The Solution has Lower Expected Passenger Time in Practice

We assume for each action, a primary delay distribution with an average of 2% of the action minimum time. This value of a is Infrabels current best estimate for morning peak hours. Similarly, [Gov07] also uses percentages between 0 to 5%.

Consider figure 1 and its caption. The left half of the figure shows the planned train time total minima and total supplements, both for the *oRiginal* timetable (R) and for the *oPtimised* timetable (P). The right half represents passenger weighted planned time for all

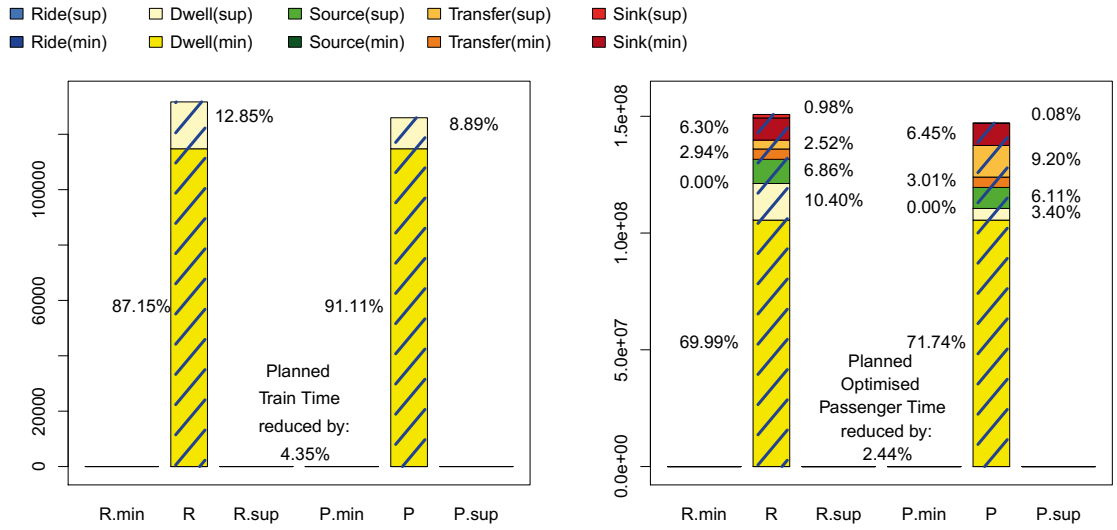


Figure 1: The *planned* time domain. The left half shows total planned train time for all trains. The right half show total passenger time for all passengers. In each box, the left bargraph shows a quantity for the *oRiginal* timetable while the right half shows the same quantity for the *oPtimised* timetable. *min* = sum of all minima, *sup* is sum of all supplements. The sum is not weighted for train time and passenger weighted for passenger time. *Source* corresponds to boarding passengers and *sink* to alighting passengers. In this *planned* domain, the shading with blue lines indicates that these actions were *summed* with ride actions.

origin-destination passenger streams with at least 50 people, again both for original and optimised timetable. There are dark and light versions of some colours (e.g.: yellow, orange). The dark colour indicates the sum of minimum times, while the lighter version indicates the sum of supplement times. The left half of figure 1 shows a decrease of total planned train supplements from 12.85% down to 8.89%. This train time supplement reduction is advantageous for the operator, since, if total train trip time now becomes less than the next lower multiple of hours, the same hourly service can be operated with one less train. [Lie07] also gave an example of this, optimising the Berlin Underground timetable.

The right hand side of figure 1 shows that the planned *passenger* weighted time reduction is a much more pronounced one, from 10.40% down to 3.40% of the same ride+dwell supplements. This is the case because they are now weighted by number of passengers.

In figure 2, instead of planned time, we show expected time, which includes primary delays and their consequences like secondary delays and missed transfers. The left half again represents train time. The right half shows passenger weighted time. The top row is the linear approximation of time as used in the optimisation model. The bottom row shows the actual non-linear time as it is evaluated post-optimisation. The same advantageous stronger supplement reduction in column 2 compared to column 1 is also present in this figure. This is the case for ride+dwell supplements but also for knock-on time. The knock-on

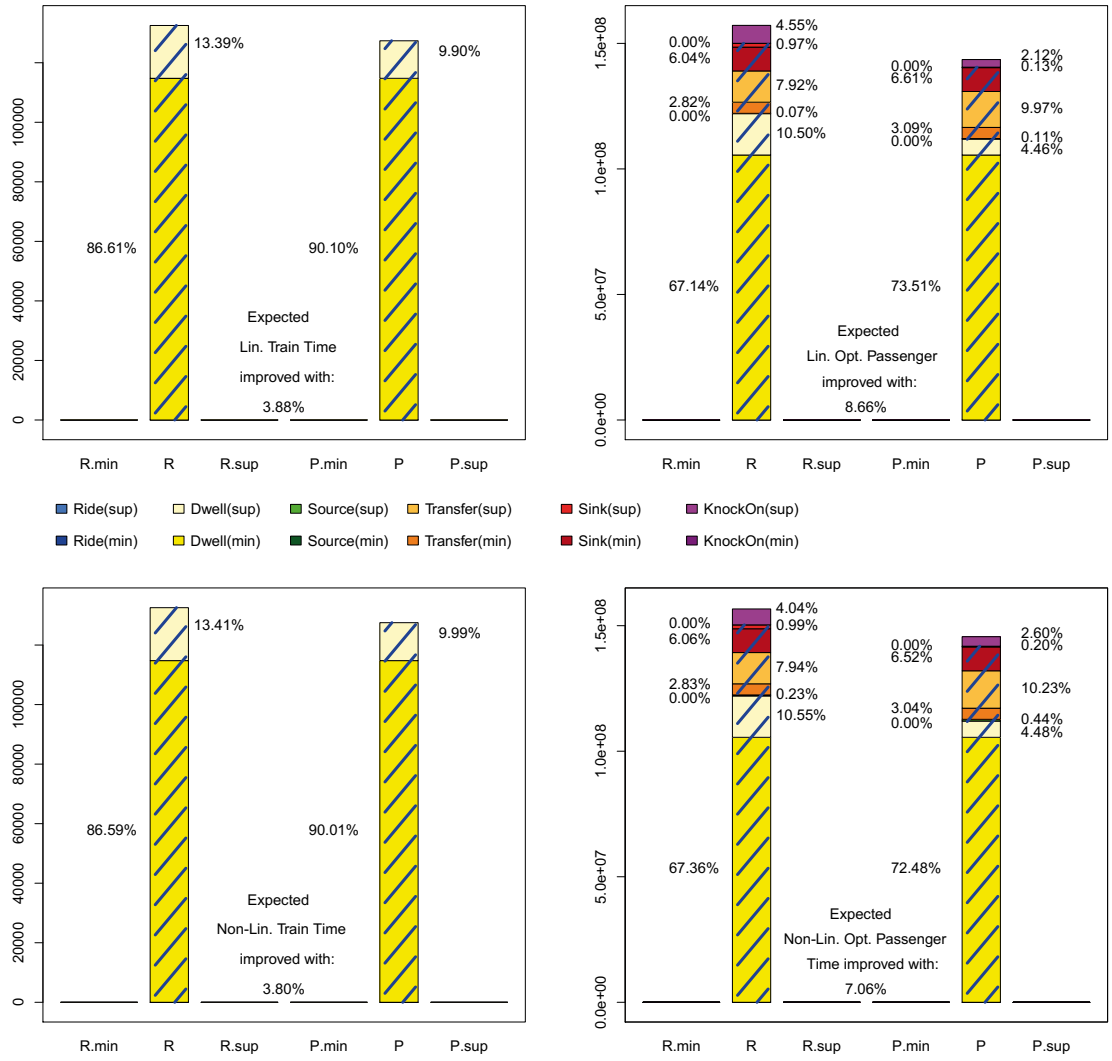


Figure 2: The *expected* time domain. Left and right are train and passenger time as in figure 1. The bottom row shows the non-linear time as used during evaluation. The top row represents the linearised approximation of it as is used during optimisation. So row 1, column 2 shows the totals achieved by optimisation of the goal function. In this *planned* domain, blue line shading indicates these actions were *convoluted* with ride actions. All figures show the case $a = 2\%$ as also reported in table 1.

component, shown as the top (purple) rectangle of the bar graphs, is reduced in percentage of the total expected passenger time from 4.55% in the original schedule to 2.12% in the optimised schedule. This is for the linearised function as used in optimisation (column 2, row 1). For the non-linearised function (column 2, row 2), post-optimisation evaluation results in a reduction from 4.04% to 2.60%. In both cases, in absolute terms, we more than halve the amount of total expected passenger knock-on delay. The solver achieves this goal by changing orders of trains on common resources and optimally choosing headway

supplements, weighing with passenger numbers and also balancing these with other goal function terms. Note that our model assumes the absence of dispatching interventions but with fewer knock-ons happening, the number of necessary dispatching interventions will be lower than in the original timetable as well.

The decrease of the ride+dwell and knock-on times is compensated only slightly by the small increase in expected transfer time. In column 2, representing evaluation on all origin-destination flows of 50 and more passengers, the total time net reduction is 8.66% (row 1, linear) and 7.06% (row 2, non-linear). The fact that the two pictures in column 2 are quite similar, demonstrates that our linearisation, even if only using 2 segments, is effective.

When we evaluate on *all* passenger streams, also the ones with fewer than 50 passengers, the result is a less grand, but still positive 0.42% reduction (non-linear). Plotting distributions of planned passenger journey time versus number of people, we saw that distributions corresponding to the major flows of column 2 are more realistic than the ones corresponding to all passenger streams. None of the major passenger flows, but a minority of the smaller ones have journey times between 2 and 3 hours for a single trip. Some of these are caused by an overenthusiastic diffusion of the zone-OD matrix to the station level [Sel11]. These travellers would most likely prefer other modes of transport. So we consider 7.06% to be our best prediction for reduction of total expected passenger time. Note that an average planned buffer of 8.89% is not enough to totally eliminate all knock-on delays, even though the assumed primary delays have only an average of 2%, seen in train time. The non-zero spread in the primary distribution explains this.

4 5

As table 1 shows, compared to the current timetable, our optimised timetables have quite some advantages. First, they respect all minimum ride- and dwell-times and all headway time buffers of 3 minutes between all train pairs on the same track section. In the original timetable sometimes minimum run times and headway times are not respected. Second, we calculated that, over all primary delay assumptions of table 1, the average chance of missing a transfer in the current timetable is at least 14.1% while in our optimised timetables it is at most 4.4%. Depending on the primary delays assumed, in our timetables the expected passenger times are between 7.06% and 0.42% lower than in the original schedule. This decrease is significant, because, of the total passenger time, the irreducible part of minimal ride and dwell times already consumes 67% in the original and 73% in the optimised timetable.

3.3 Computation Speed: The Solution is Returned Quickly

Using the solver abstraction part of the software library milp-logic [Sel12], which we developed and open sourced, as shown in table 1, Gurobi 5.5.0 was able to return a solution for the whole train set, for any primary delay distributions assumed, within about one hour. This is a big improvement compared to the current manual timetabling process that takes many

[4] All periodic trains are repeated every hour. The P trains occur just once in practice, but due to our cyclic timetable, time slots are automatically reserved for them in every non-peak hour too.

[5] Outside the peak period, these empty P slots can be used for freight trains if desirable.

Table 1: Increasing primary delays, characterised by their average of $a\%$ of minimum dwell and ride times. The first column shows $a\%$. Column 2 and 3 show the computation time and the MILP gap achieved. We ran Gurobi 5.5.0 on an Apple MacBook Pro with 2.6GHz Intel i7 processor and 16GB 1.6GHz DDR3 memory. For the first set of result rows, the gap desired was set slightly above what was obtained as the gap of the first returned solution in earlier trials. The results in the last row are obtained by reduction of the desired gap by 1% compared to the first row. Graph size: 203 hourly trains, 5355 ride, 5152 dwell, 17553 major transfer, 31696 knock-on and 166 turn-around edges. Model size: 42609 supplement decision variables, 49415 integer decision variables, 41128 goal function terms for major flows and 58441 evaluation function terms for all flows.

a	solver time	MILP gap	major flows linearised ko-time reduction	major flows linearised time reduction	major flows non-linearised time reduction	all flows linearised time reduction	all flows non-linearised time reduction	missed transfer probability orig.	missed transfer probability opt.
%	min.	%	%	%	%	%	%	%	%
2	95	76.2	57	8.66	7.06	1.71	0.42	14.1	2.2
4	43	71.0	52	6.61	4.42	0.84	-1.41	14.6	4.2
6	75	61.3	63	7.65	5.73	2.07	0.13	15.1	1.8
8	66	61.3	59	5.83	3.86	0.40	-1.61	15.6	4.4
2	112	72.6	66	10.58	9.19	2.54	1.31	14.1	2.6

human planners many months.

4 Conclusions and Further Work

This paper has three main contributions. Firstly, we analytically derived the expected passenger time experienced due to knock-on delays as a function of (i) the headway minima, (ii) the chosen headway supplements in a timetable and (iii) expected train delays and linearised this function, so that it can be used for linear optimisation. Secondly, we used the linearised functions as a method to minimise secondary delays, together with other expected passenger time, in a system containing all hourly trains in Belgium. Our results show that we can more than halve the amount of expected passenger knock-on delay in practice. Also, even with addition of many terms to the goal function, optimisation times for the Belgian timetable are only about one hour. Supposing primary delay distributions with an average of 2% of the minimal time of their corresponding actions, our improved timetable reduced expected passenger time for realistic passenger streams by 7.06% compared to the current one. Finally, although restricting the search space and using curtailed goal functions are the easy way to try to reduce solver time, we show that defining an all-encompassing goal function and searching the full solution space can lead to more desirable results: guaranteed feasibility, optimality and even lower solver times.

As for further work, we want to reduce our MILP gap, refine our minimum transfer time differentiating it by station and calibrate our primary delay distributions with train and location specific delays measured in practice.

References

[Dew11] T. DEWILDE, P. SELS, D. CATTRYSSE, and P. VANSTEENWEGEN: "Defining Robustness of a Railway Timetable". In: *Proceedings of 4th Int'l Seminar on Railway Operations Modelling and Analysis (IAROR)* (Feb. 16–18, 2011).

[Gov07] R. M. GOVERDE: "Railway timetable stability analysis using max plus algebra". In: *Transportation Research Part B: Methodological* 41 (2007), pp. 179–201.

[Gov10] R. M. GOVERDE: "A delay propagation algorithm for large-scale railway traffic networks". In: *Transportation Research Part C* 18 (2010), pp. 269–287.

[Han08] I. A. HANSEN and J. PACHL: *Railway Timetable and Traffic: Analysis, Modelling, Simulation*. Vol. 1. Eurailpress, Hamburg, Germany, Jan. 2008, pp. 1–288.

[Kro09] L. KROON, D. HUISMAN, E. ABBINK, P.-J. FIOOLE, M. FISCHETTI, G. MARÓTI, A. SCHRIJVER, and R. YBEMA: "The New Dutch Timetable: The OR Revolution". In: *Interfaces* 39 (2009), pp. 6–17.

[Lie07] C. LIEBCHEN: "Periodic Timetable Optimization in Public Transport". In: *Operations Research Proceedings* 2006 (2007), pp. 29–36.

[Mus13] L. MUSSONE and R. WOLFLER CALVO: "An analytical approach to calculate the capacity of a railway system". In: *EJOR* 228 (2013), pp. 11–23.

[Sel11] P. SELS, T. DEWILDE, D. CATTRYSSE, and P. VANSTEENWEGEN: "Deriving all Passenger Flows in a Railway Network from Ticket Sales Data". In: *Proceedings of 4th Int'l Seminar on Railway Operations Modelling and Analysis (IAROR): RailRome2011* (Feb. 16–18, 2011). URL: `4c4u.com/RR2011.pdf`.

[Sel12] P. SELS: *milp-logic: a C++ MILP Solver Abstraction Layer with a C++ Boolean Modelling Layer on Top.* Sept. 2012. URL: `code.google.com/p/milp-logic`.

[Sel13a] P. SELS, T. DEWILDE, D. CATTRYSSE, and P. VANSTEENWEGEN: "An Optimal Timetable for the Whole Belgian Railway Network". 2013.

[Sel13b] P. SELS, T. DEWILDE, D. CATTRYSSE, and P. VANSTEENWEGEN: "Expected Passenger Travel Time as Objective Function for Train Schedule Optimization". In: *Proceedings of 5th International Seminar on Railway Operations Modelling and Analysis (IAROR): RailCopenhagen2013* (May 13–15, 2013). URL: `4c4u.com/RC2013.pdf`.

[Spa13] D. SPARING, R. M. GOVERDE, and I. A. HANSEN: "An Optimization Model for Simultaneous Periodic Timetable Generation and Stability Analysis". In: *Proceedings of 5th International Seminar on Railway Operations Modelling and Analysis (IAROR): RailCopenhagen2013* (May 13–15, 2013).

Corresponding author: Peter Sels, KU Leuven, University of Leuven, Centre for Industrial Management/Traffic & Infrastructure, 3001 Leuven, Belgium, phone: +32 486 95 67 97, e-mail: sels.peter@gmail.com

Capacity-Utilized Integration and Optimization of Rail Freight Train Paths into 24 Hours Timetables

Peter Großmann, Alexander Labinsky, Jens Opitz, Reyk Weiß
Technische Universität Dresden

Abstract

Timetables for railway networks are still manually constructed and in most cases not optimized. Besides the fact that scheduling software for periodic timetabling exists, in reality both periodic and non-periodic train paths have to be included into operating timetables. Therefore, we propose a new model based on the Periodic Event Scheduling Problem for integrating and optimizing all, non-periodic train paths into periodic timetables simultaneously, utilizing the capacity of the network. The corresponding LP model will be introduced and discussed in detail which reflects all constraints for efficient and conflict-free freight train path timetabling.

Keywords: rail transport, rail freight transport, scheduling, timetabling, 24 hours timetables, optimization, Periodic Event Scheduling Problem, PESP

1 Introduction

Railway networks must be designed and operated intelligently to ensure conflict-free and optimized railway traffic. Every train movement between two stops has to be scheduled in advance guaranteeing a safe journey without conflicts with other train movements on the same track [Pac08]. Therefore, the construction of timetables is a core part of railway operations.

This paper gives insight in a new project currently under development at the Chair of Traffic Flow Science at TU Dresden in close collaboration with DB Netz AG, which allows the construction of automatically optimized rail freight train paths into an already existing passenger rail timetable for 24 hours periods of time as depicted in Fig. 1. For the first time, this allows the direct transfer of the timetable into daily operations and introduces automatically constructed timetables ready to use in real operations. Additionally, these timetables will be generated much faster and thus, offer quick comparisons of different

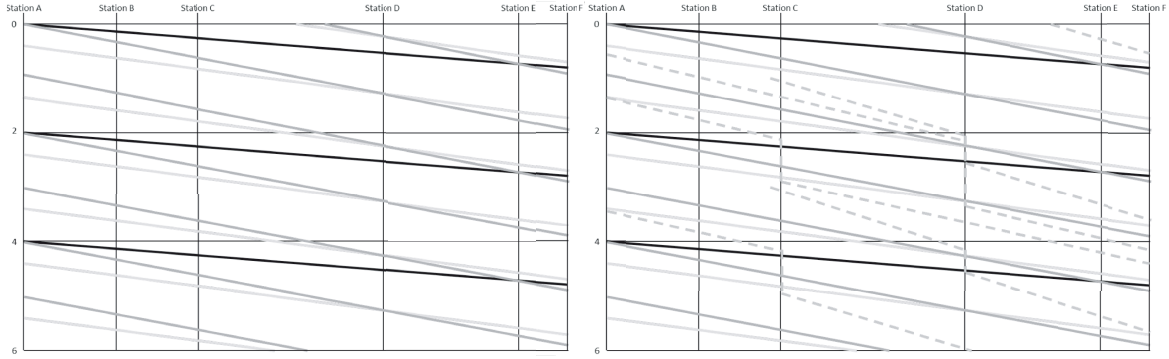


Figure 1: Cutout of a service planning diagram for a periodic timetable without additional non-periodic train paths (left) and with additional non-periodic train paths (right)

scenarios with altered constraints. For this aim, we formulate a cubic objective function with linear constraints. This objective function has to be linearized while all boundary conditions inherent to railway traffic have to be met.

The following section explains the realization of today's railway timetabling. Section 3 describes the insertion of non-periodic trains into existing timetables, whereas in Subsection 3.1 the boundary conditions needed for the model to work are identified, while in Subsection 3.2 the mathematical model is introduced. In Section 4 we present results obtained by case studies for a theoretical test case and a heavy-loaded railway section in Southern Germany. Section 5 concludes our findings and gives an outlook on further investigations.

2 Railway Timetabling

Scheduling trains is a very complex task which until today still needs experienced specialists to construct reliable timetables for heavily loaded railroads [Pac08]. Finding new train paths in an already crowded timetable is a time-consuming, iterative process which often results in lower quality, for example unnecessary long travel times for individual train paths. Today most rail companies use computer software for timetabling. However, is more related to a virtual drawing board than an automatic scheduling program (for example RUT-K as used by Deutsche Bahn AG). Hence, timetables are still mainly generated by manual effort – and therefore, not mathematically optimized – as they were in the 19th century.

On the other hand, the Periodic Event Scheduling Problem (PESP) as introduced by Serafini and Ukovich [Ser89], which is known to be NP-complete, is the origin of a new approach to schedule periodic timetables automatically [Nac98].

A Periodic Event Network (PEN) is defined by a network that consists of a set of nodes $\mathcal{N}$ and a set of edges $\mathcal{E}$ whose events are repeated by the period t_T and every edge $a \in \mathcal{E}$ provided with a modulo time span $[t_{\min,a}, t_{\max,a}]_{t_T}$ defined by lower and upper bound $\vec{t}_{\min}$ and $\vec{t}_{\max}$, respectively. A PEN is feasible if and only if for all $a = (i, j) \in \mathcal{E}$

$$\vec{t}_{\min,a} \leq T_j - T_i - t_T \cdot z_a \leq \vec{t}_{\max,a} \tag{1}$$

with T_i point of time for event i and z_a integer modulo parameter of t_T. All T_i together form a timetable for the PEN. PESP is the decision problem whether a valid timetable for a PEN exists [Ser89].

This allows the development of highly advanced scheduling software like TAKT which is able to generate optimized timetables for complex networks in hours compared to weeks needed even by experienced specialists [Opi09]. Nevertheless, these programs still are not able to cope with non-periodic trains and changing frequencies during the day. Furthermore, they cover only parts of a given day like reflecting peak hours and off-peak periods. Therefore, timetables generated with these programs are not suited for daily operations as non-periodic trains and train lines changing frequencies during the day are common problems in timetabling [Pac08].

3 Non-periodic Train Path Insertion into Existing Timetables

The current project focuses on rail freight train paths as non-periodic train paths and works with fixed rail passenger train paths. This limitation has several practical reasons: Firstly, rail passenger services in Germany are often fixed by long-term contracts with political bodies and therefore inelastic; secondly, rail passenger services are mostly periodic and can already be solved by state-of-the-art software as shown above; thirdly, non-periodic train paths as needed for rail freight trains are currently constructed manually, therefore non-optimized and thus, need most attention.

Nevertheless, the insertion of rail freight train paths will serve as an initial, important step towards fully automatized timetabling in the future, optimizing both periodic and non-periodic train paths simultaneously.

3.1 Boundary Conditions

In order to generate optimized non-periodic train paths, several boundary conditions have to be fulfilled. First of all, we need a given timetable with fixed rail passenger services. This timetable works as a frame for non-periodic train paths, since these cannot use tracks which are already used by passenger train paths.

After the initial frame of rail passenger train paths is loaded and fixed, the operator specifies the number of necessary new non-periodic train paths. While these train paths are expected to have a requested order of departure, it is mandatory that routing is chosen freely by the algorithm. Hence, infrastructure data of all usable railroad tracks has to be available. On behalf of the given data, all possible train paths are split into a finite set of sub paths, so called InfraAtoms, which are discretized and used to construct the new non-periodic train paths, always reflecting actual driving dynamics. This approach already showed remarkable results [Wei12].

The way day change-overs are harmonized must be regarded as well. A timetable is only useful if it seamlessly adjoins to its preceding and following timetables. For this purpose, the

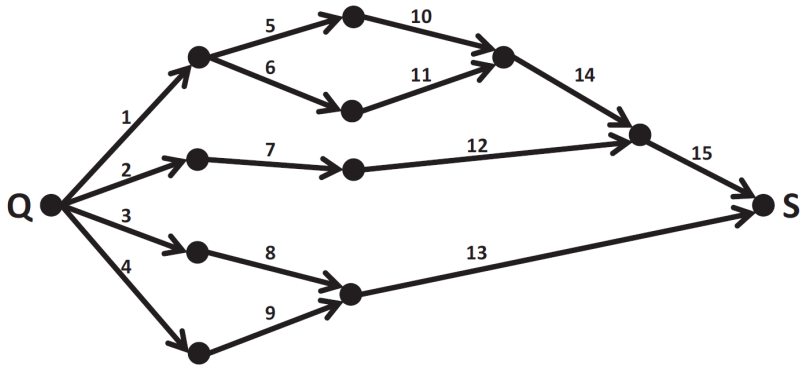


Figure 2: Example network with a set of 15 InfraAtoms $\mathcal{N} = \{1, ..., 15\}$ including 4 source InfraAtoms $O(Q) = \{1, 2, 3, 4\}$ and 2 sink InfraAtoms $I(S) = \{13, 15\}$. *Source: [Wei12]*

loaded timetable is longer than an actual day to allow the construction of rail freight train paths which affect the next 24 hour period. These train paths will be copied into the fixed timetable for the next day and cannot be altered to allow the construction of a consecutive chain of timetables, for example a week or even longer periods.

3.2 Insertion Model of Optimized Non-periodic Train Paths

Our goal is scheduling a timetable with reasonable quality. Therefore, our model aims to minimize the measurement of quality for single rail freight train paths, the parameter BFQ which is defined as:

$$\sum_i \xi_i = \sum_i \frac{t_i}{t_{\min}} \tag{2}$$

with ξ_i as BFQ of train path i, t_i as total runtime of the current train path, and $t_{\min}$ as fastest runtime possible for a train path.

For our model we define the entirety of all InfraAtoms as $\mathcal{A}$, further mapping the subset of InfraAtoms going into a given node n (or set of nodes) as $I(n)$, whereas the subset of outgoing InfraAtoms of a given node n (or set of nodes) is mapped as $O(n)$. The set of all nodes is defined as $\mathcal{N}$ while the subset of all starting nodes is defined as Q, whereas the subset of all destination nodes is defined as S. Figure 2 visualizes such a graph by example. Since every InfraAtom has only one periodic event [Wei12], nodes and InfraAtoms are handled equivalently.

Additionally, the number of train paths to be newly inserted is defined as τ and all already existing train paths are defined as $\mathcal{P}$.

Furthermore, we need the binary variable $x_{i,j}$, which shows whether the InfraAtom j is used by train path i ($x_{i,j} = 1$) or not ($x_{i,j} = 0$). The travel time for $x_{i,j}$ is defined as $t_{i,j}$ and the departure time of train i at the beginning of InfraAtom j is defined as $T_{i,j}$. Halting times at the end of InfraAtoms are defined as T_H and headways between two trains i and l are defined as $t_{head;i,l}$.

We further need a $\mathcal{C} \gg 0$ and a binary variable $z_{i,j;l,k}$, indicating whether InfraAtom l used by train path k follows on InfraAtom j used by train path i ($z_{i,j;l,k} = 1$) or not ($z_{i,j;l,k} = 0$). We additionally state that for any $y \in \{0,1\} : y + \overline{y} = 1$ and use this relation for binary variables accordingly.

Finally, we introduce a variable φ, which defines the maximum allowed permutation for a given event. By a smart choice of φ, we are able to reduce the number of constraints of the model substantially, since this results in a complexity of $\mathcal{O}(\varphi\tau)$ in contrast to a complexity of $\mathcal{O}(\tau^2)$ without φ.

With the definitions given above, we formulate our objective function and accompanying constraints as follows:

$$\sum_{i\in\{1,\dots,\tau\}} \sum_{j=(n_1,n_2)\in\mathcal{A}} (t_{i,j} \cdot x_{i,j} + \sum_{k\in O(n_2)} x_{i,j} \cdot x_{i,k} \cdot (T_{i,k} - T_{i,j} - t_{i,j})) \to \min \tag{3}$$

s.t.

$\forall i \in \{1,\dots,\tau\}$:

$$\sum_{j\in Q} x_{i,j} = \sum_{j\in S} x_{i,j} = 1, \tag{4}$$

$\forall i \in \{1,\dots,\tau\}\ \forall n \in \mathcal{N} \setminus (Q \cup S)$:

$$\sum_{j\in I(n)} x_{i,j} - \sum_{j\in O(n)} x_{i,j} = 0, \tag{5}$$

$\forall i \in \{1,\dots,\tau\}\ \forall n \in Q\ \forall j \in O(n)$:

$$T_{i,j;\min} \le T_{i,j} \le T_{i,j;\max}, \tag{6}$$

$\forall i \in \{1,\dots,\tau\}\ \forall n \in Q\ \forall j,k \in O(n)$:

$$T_{i+1,j} - T_{i,k} \ge 0, \tag{7}$$

$\forall i \in \{1,\dots,\tau\}\ \forall j = (n_1,n_2) \in \mathcal{A}\ \forall k \in O(n_2)$:

$$T_{i,k} - T_{i,j} + \mathcal{C} \cdot \overline{x_{i,j}} + \mathcal{C} \cdot \overline{x_{i,k}} \ge t_{i,j} + t_{H,\min}, \tag{8}$$

$$T_{i,k} - T_{i,j} - \mathcal{C} \cdot \overline{x_{i,j}} - \mathcal{C} \cdot \overline{x_{i,k}} \le t_{i,j} + t_{H,\max}, \tag{9}$$

$\forall j,k \in \mathcal{A}\ \forall m \in \{1,\dots,\varphi\}\ \forall i \in \{1,\dots,\tau-m\}$; *j, k share blocks*:

$$T_{i+m,k} - T_{i,j} + \mathcal{C} \cdot \overline{z_{i,j;i+m,k}} + \mathcal{C} \cdot \overline{x_{i,j}} + \mathcal{C} \cdot \overline{x_{i+m,k}} \ge t_{\mathit{head};i,i+m;1}, \tag{10}$$

$$T_{i+m,k} - T_{i.j} - \mathcal{C} \cdot z_{i,j;i+m,k} - \mathcal{C} \cdot \overline{x_{i,j}} - \mathcal{C} \cdot \overline{x_{i+m,k}} \le -t_{\mathit{head};i,i+m;2}, \tag{11}$$

$\forall i \in \{1,\dots,\tau-(\varphi+1)\}\ \forall j,k \in \mathcal{A}$; *j, k share blocks*:

$$T_{i+\varphi+1,k} - T_{i,j} + \mathcal{C} \cdot \overline{x_{i,j}} + \mathcal{C} \cdot \overline{x_{i+\varphi+1,k}} \ge t_{\mathit{head};i,i+\varphi+1;\min} \tag{12}$$

$\forall p \in \mathcal{P}\ \forall i \in \{1,\dots,\tau\}\ \forall j \in \mathcal{A}$; *j, p share blocks*:

$$T_{i,j} + \mathcal{C} \cdot \overline{z_{i,j;p}} + \mathcal{C} \cdot \overline{x_{i,j}} \ge t_{\mathit{head};i,p;1} + T_p, \tag{13}$$

$$T_{i,j} - \mathcal{C} \cdot z_{i,j;p} - \mathcal{C} \cdot \overline{x_{i,j}} \le -t_{\mathit{head};i,p;2} + T_p. \tag{14}$$

$\forall i \in \{1,\dots,\tau\}\ \forall l \in \{i,\dots,i+\varphi\}\ \forall j,k \in \mathcal{A}$: $T_{i,j} \in \mathbb{Z},\ x_{i,j}, z_{i,j;l,k} \in \{0,1\}$

The objective function (3) minimizes the sum of the travel time for all used InfraAtoms by

all new train paths. Additionally, the halting times for all $j \rightarrow k$ which include a halt at the end of InfraAtom j have to be added up and minimized as well. For this reason, our model has a cubic objective function with linear constraints. For convenience, we did not note the linearized form of (3) as this would have resulted in more complex constraints.

The following constraints have to be formulated: As shown in (4), every given train path i has only one starting node and only one destination node. For all other nodes on i, for every ingoing InfraAtom must also exist an outgoing InfraAtom (see (5)). Furthermore, we state in (6) that the departure of i at the starting node has to be between the minimum and maximum departure time specified in advance and in (7) that it has to be scheduled before the departure of the following train path $i+1$. Therefore, we need the fixed number of train paths τ to ensure a feasible model.

The travel time for a given train path i which is halting at the end of InfraAtom j has to be set between the sum of the travel time for j and the minimum halting time (as shown in (8)) and the sum of the travel time for j and the maximum halting time (see (9)).

Of most interest are blocking times, as these ensure that two InfraAtoms j and k do not use the same block at the same time guaranteeing conflict-free train paths. Hence, we formulate constraint (10) indicating if $z_{i,j;i+m,k} = 1$ is given, the specified headway for both InfraAtoms $t_{head;i,i+m;1}$ has to be less than the difference of $T_{i,j}$ and $T_{i+m,k}$. This holds as well if both trains change sequence. In this case, as shown in (11), the headway has to be less than the difference of $T_{i,j}$ and $T_{i+m,k}$. To further simplify the problem, we also state that permutation of train paths at a given node, or event, is limited to φ. Therefore, we can propose in (12) that the headway between train paths i and $i+\varphi+1$ always has to be greater than minimum the headway $t_{head;i,i+\varphi+1;min}$.

At last, we have to ensure that blocking times already used by existing train paths must not be violated. As these train paths p are fixed, in (13), we set $T_{i,j}$ greater than the sum of the specified headway and T_p if train path i follows on p and in (14) less than the difference of T_p and specified headway if p follows on i depending on the sequence $z_{i,j;p}$.

Based on the fact that the model is discrete (all variables are binary or integer), the whole model is encoded into a propositional formula to formulate a Boolean satisfiability problem (SAT) [Bie09]. This transformation is not part of this paper and will be outlined in a subsequent publication.

With the help of state-of-the-art SAT solvers, an initial assignment for the proposed problem can be found [Gro12]. This initial solution can be further optimized by using state-of-the-art LP solvers.

4 Computational Results

In order to prove feasibility of the presented model, we calculated both a theoretical test case and a real-world instance. Figure 3 shows a timetable obtained for the theoretical test case. The goal was to include 18 new rail freight train paths (dark grey) with two different classes of maximum speed into an already existing five hours frame of rail passenger train paths,

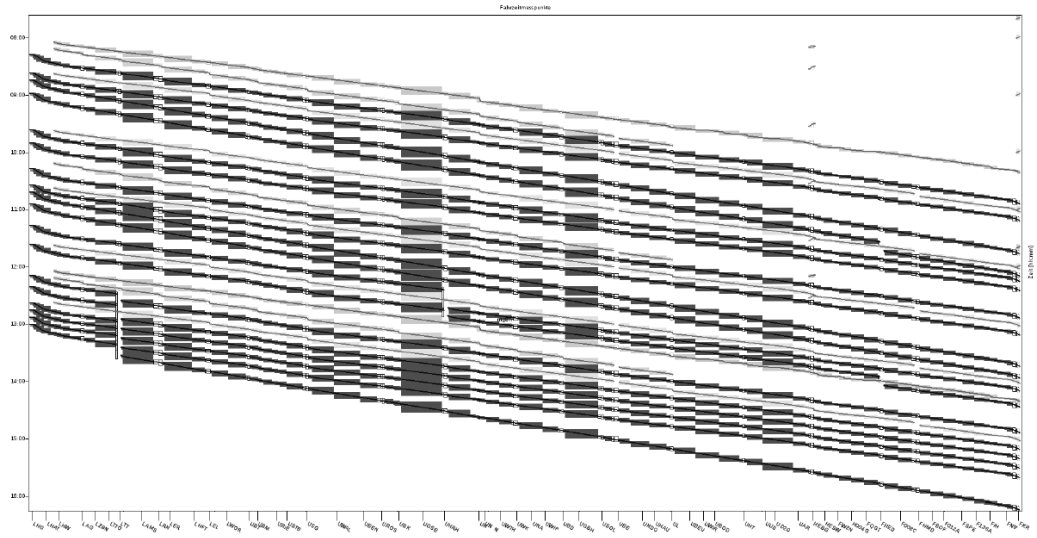


Figure 3: Timetable for test case with 18 rail freight train paths (dark grey) included into existing rail passenger timetable

which proved successful. We could include all new train paths into the required departure time spans. As for quality, we obtain an average BFQ as described in (2) of 1.076 and a maximum BFQ of 1.162 in about half an hour. The optimization process is stopped when reaching a sufficient level of quality which impedes to give an exact amount of time needed to fully solve the problem.

Based on these results, we solved another instance, this time for the real-world railway line between Mannheim and Basel, which is part of the heavily loaded railway network along the river Rhine. Our goal was to include 90 new rail freight train paths during a period of 24 hours, which proved successful as well. All required new train paths could be included into the defined departure time spans. The average quality as expressed by BFQ is 1.311, with a maximum BFQ for a single train path of 1.605. Overall calculation time is approximately 6 hours.

This findings correspond to the remarkable results already obtained by previously used methods for automated timetabling [Nac98; Opi09; Gro12].

5 Conclusion

In this paper, we showed a new approach for automatic scheduling of optimized rail freight train paths. We modified PESP to allow the insertion of non-periodic trains into an already existing 24 hours timetable. This allows the use of these timetables for daily operations, due to the fact that periodic and non-periodic train paths will be constructed conflict-free.

Consequently, we are able to optimize railway traffic and schedule timetables much faster than today. This allows a more intelligent use of the existing infrastructure as the specialists currently needed for timetabling are able to invest more time in investigating alternative

timetable scenarios. On the other hand, customers get faster and more optimized train paths as currently sold. Therefore, railway operation will be hopefully more efficient than today.

Of course, the findings as presented in this paper can only be considered as a first step in a chain of further investigation. The future goal has to be a fully automatic constructed timetable which includes both periodic and non-periodic train paths. Nevertheless, our findings are an important step for intelligent railway scheduling as we showed that the insertion of high-quality railway freight train paths into existing railway passenger timetables is possible in reasonable computing times.

References

[Bie09] A. BIERE, M. HEULE, H. van MAAREN, and T. WALSH: *Handbook of Satisfiability*. IOS Press, 2009.

[Gro12] P. GROSSMANN, S. HÖLLDOBLER, N. MANTHEY, K. NACHTIGALL, J. OPITZ, and P. STEINKE: "Solving Periodic Event Scheduling Problems with SAT". In: *IEA/AIE*. Vol. 7345. LNAI. Springer, 2012, pp. 166–175.

[Nac98] K. NACHTIGALL: "Periodic Network Optimization and Fixed Interval Timetable". Habilitation thesis. University Hildesheim, 1998.

[Opi09] J. OPITZ and P. NACHTIGALL: *Automatische Erzeugung und Optimierung von Taktfahrplänen in Schienenverkehrsnetzen*. Logistik, Mobilität und Verkehr. Gabler Verlag, 2009. ISBN: 9783834921284. URL: http://books.google.de/books?id=PULo7htW77UC.

[Pac08] J. PACHL: *Systemtechnik des Schienenverkehrs*. 5th. Wiesbaden: Vieweg + Teubner, 2008, p. 285. ISBN: 978-3834814289.

[Ser89] P. SERAFINI and W. UKOVICH: "A Mathematical Model for Periodic Scheduling Problems". In: *SIAM J. Discrete Math.* 2.4 (1989), pp. 550–581.

[Wei12] R. WEISS, J. OPITZ, and K. NACHTIGALL: "A novel approach to strategic planning of rail freight transport". In: *OR Hannover, Presentation*. 2012.

Corresponding author: Peter Großmann, Technische Universität Dresden, "Friedrich List" Faculty of Transportation and Traffic Sciences, Chair of Traffic Flow Science, 01062 Dresden, Germany, phone: +49 351 463 36506, e-mail: peter.grossmann@tu-dresden.de

The State-of-the-art Realization of Automatic Railway Timetable Computation

Michael Kümmling, Peter Großmann, Karl Nachtigall, Jens Opitz, Reyk Weiß
Technische Universität Dresden

Abstract

We describe the Periodic Event Scheduling Problem (PESP) including extensions like symmetry and its power, solving it using SAT solvers and optimizing timetables as implemented in the software system TAKT. PESP still lacks efficient applications for large-scale intermeshed railway networks. Thus, we discuss several strategies for efficiently resolving conflicts and intelligently splitting timetable problems. First applications give promising results.

Keywords: intelligent timetabling, timetable optimization, PESP, timetabling, SAT

1 Introduction

Today, planning and disposition of railway transport is often manual labor – computers only hold, manage and visualize data. These labor-intensive processes lead to mostly evolutionary timetabling, such that every year only the necessary changes are done to fit the timetable to changed infrastructure or changed route network. Using state-of-the-art techniques, it is nearly impossible to build only one fully optimized railway timetable from scratch for large and intermeshed railway networks like the German one. As this obviously does not tap the railway's full potential, diverse high-quality timetable variants have also a key role in intelligent design of tomorrow's railway infrastructure and in reliable stability and capacity analysis leading to efficient and sound railway networks.

The key to automation of railway timetabling and real-time rescheduling is the utilization of proper models and algorithms. As railways are a quite long-standing business, they have grown large bundles of complicated operational rules and versatile constraints. Hence, a model is needed which covers a wide range of possible constraints.

The Chair of Traffic Flow Science at TU Dresden develops since several years in close collaboration with DB Netz AG the software system TAKT that tackles exactly these kind of issues and solves them by state-of-the-art operations research techniques. The connection and interaction between the different approaches will be subject of this work.

In Section 2 we give the preliminaries for periodic event scheduling. After showing the possibilities for solving timetabling instances in Section 3, we introduce in Section 4 the state-of-the-art conflict resolving of infeasible instances. Presenting our computational results in Section 5, we conclude the work in Section 6 and give a further scientific outlook.

2 Periodic Event Scheduling Problem

In the last 15 years, the Periodic Event Scheduling Problem (PESP) established as one of the most suitable problem formulations for periodic timetabling. It is introduced by Serafini and Ukovich [Ser89]. The related periodic event network permits flexible representation of almost all periodic timetable's constraints. For instance, PESP and its implications are discussed in detail by Nachtigall [Nac98] and Opitz [Opi09].

The operating program, which is the base of the timetabling problem, contains routes $\mathcal{L}$ running on a railway network with stations $\mathcal{S}$. Each Route $L \in \mathcal{L}$ serves a specified sequence of Stations $S \in \mathcal{S}$. All constraints are modeled into an event network. Its nodes in $\mathcal{V}$ represent arrival events $(L, arr, S) \in \mathcal{V}$ and departure events $(L, dep, S) \in \mathcal{V}$. The schedule $\vec{T} \in \mathbb{Z}^{|V|}$ assigns to every event $i \in \mathcal{V}$ a potential $T_i \in \mathbb{Z}$, $0 \leq T_i < t_T$. In a periodic timetable with period $t_T \in \mathbb{N}^+$ the event happens periodically at all times $T_i + zt_T$, $z \in \mathbb{Z}$.

The network's arcs $a \in \mathcal{A}\colon i \to j$ are basically time consuming processes. All arcs' processing times are constrained by lower bounds $t_{min,a}$ and upper bounds $t_{max,a}$. This range is also written as $[t_{min,a}, t_{max,a}]_{t_T}$. A timetable $\vec{T}$ is considered valid if and only if

$$\forall a \in \mathcal{A}\colon \exists z_a \in \mathbb{Z}\colon t_{min,a} \leq T_j - T_i - z_a t_T \leq t_{max,a}. \tag{1}$$

The *lower slack* y_a is the deviance of the actual processing time from the lower bound such that

$$0 \leq y_a = T_j - T_i - z_a t_T - t_{min,a} < t_T. \tag{2}$$

The periodic event scheduling problem (PESP) is the decision, whether there exists any valid timetable for a given periodic event network $\mathcal{N} = (\mathcal{V}, \mathcal{A}, \vec{T})$. For feasible problems, a timetable can be calculated.

This universal model allows the modeling of running times, dwell times, headways and transfer times. For instance, trains are encoded as alternating sequences of running activities $(L, dep, S) \to (L, arr, S')$ and stops $(L, arr, S) \to (L, dep, S)$. Headways between different routes include both safety headways representing the permitted minimum headway and also evenly distributed headways of different trains running partly on the same railway line. Transfer times include several of different requirements: vehicle transfers, staff transfers and passenger transfers.

Likewise, symmetry is a common requirement in periodic timetabling. A route and its associated returning route are considered to be symmetric, if their arrival and departure times are aligned symmetrically along symmetry axis in time (see Eq. (3)), which is called symmetry minute s and is equal in the whole network. Hence, the trains meet themselves at point of

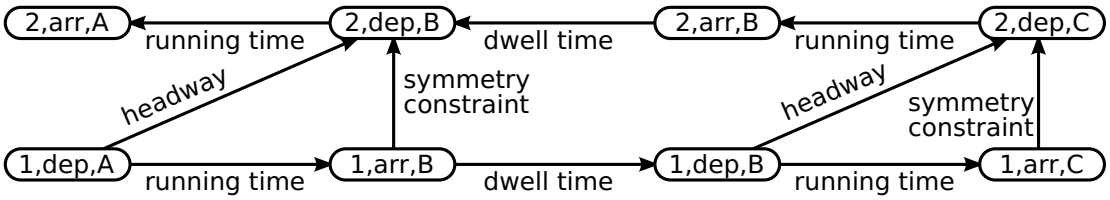


Figure 1: An event network of two routes on a single track line with stations A, B and C

time s. The symmetric timetables' advantage is the satisfied symmetric property of transfers. Thus, all transfers automatically are fulfilled in both directions. This constraint is special, as it cannot be modeled by Eq. (1) as it was proven by Liebchen [Lie06]. Subsequently, the model has to be extended by additional constraints $a \in \mathcal{A}_S \colon i \to j$, where i is the arrival event of one route and j the departure event of the associated returning route. Exact symmetry for every routes' stops would often result in too restricted problems due to infrastructural restrictions and slightly different running times. Therefore, a certain maximum absolute deviation from symmetry $d_a \in \mathbb{N}$ is permitted. The actual symmetry deviation is denoted as symmetry slack $y_a \in \mathbb{Z}$. Applying the permitted slack (5) to formulation of symmetry axis (3) results in an inequation (6) quite similar to (1).

$$T_j - (s + y_a) - z_a t_T = (s + y_a) - T_i \tag{3}$$

$$T_i + T_j - z_a t_T = 2s + 2y_a \tag{4}$$

$$-d_a \leq y_a \leq d_a \tag{5}$$

$$2s - 2d_a \leq T_i + T_j - z_a t_T \leq 2s + 2d_a \tag{6}$$

Thus, in extension of requirement (1) a timetable is only considered valid if it holds as well:

$$\forall a \in \mathcal{A}_S \colon \exists z_a \in \mathbb{Z} \colon 2s - 2d_a \leq T_i + T_j - z_a t_T \leq 2s + 2d_a \tag{7}$$

The PESP extended by symmetry constraints allows fully modeling the standard integrated timetables [Opi09]. Figure 1 shows a simple example PESP network with symmetry constraints.

3 Solving PESP in TAKT

PESP is proven to be NP-complete [Ser89]. Hence, solving real-world PESP instances is a challenging task [Opi09]. The currently most efficient approach solving PESP is conducted by using state-of-the-art SAT solvers [Gro11]. SAT is the boolean satisfiability problem determining if there exists any interpretation satisfying a given propositional formula. SAT is likewise NP-complete [Coo71], yet, for SAT very efficient solvers exist [Man10]. It was shown

that SAT solvers outperform all previously known approaches for solving PESP despite the additional time needed for encoding and decoding SAT instances [Gro12b].

Propositional logic uses boolean variables $p \in \mathcal{R}$. Literals L are either variables p or their negation $\neg p$. Clauses are disjunctions of literals $c = \bigvee_i L_i$. Propositional formulas in conjunctive normal form (CNF) are conjunctions of clauses $\mathcal{F} = \bigwedge_j c_j$. An interpretation J assigns to every variable either the value true or false, denoted as t or f, respectively. An formula $\mathcal{F}$ is satisfiable ($\mathcal{F}^J = t$) if and only if there exists a J such that all clauses contain at least one literal assigned to t under J.

The periodic event network's constraints are encoded to propositional formulas using order encoding which is introduced for general finite ordered domains [Tan11]. All potentials T_n are encoded to $t_T - 1$ boolean variables $p_{n,i}$, whereas $p_{n,i} = t$ represents $T_n \leq i$ and thus, $p_{n,i} = f$ represents $T_n > i$. Subsequently, all constraints can easily be encoded to clauses by excluding for each constraint all invalid combinations of values (T_i, T_j). This results in the equivalence of searching for an interpretation J satisfying $\mathcal{F}$ and searching for a valid timetable $\vec{T}$ for periodic event network $\mathcal{N}$. For further reading on encoding PESP to SAT we refer to the literature [Gro12b].

An interpretation J satisfying $\mathcal{F}$ can be easily decoded to a timetable $\vec{T}$ by reversing the described encoding. Solving PESP by this approach results in one valid timetable, since it is a decision problem. Global timetable optimization can be achieved by minimizing the weighted sum of slacks using integer linear programming (ILP). Lots of different objectives can be modeled by weighting factors, for example sum of journey time for all passengers or the number of needed train sets [Opi09].

4 Resolving Conflicts

Although, solving PESP is a challenging task, usually only solving the initially formulated timetable problem is not the scope of work as almost all real timetable problems initially are not satisfiable. This is reasoned by the fact that at first the constraints are arranged idealistically tight, for example dwell times are set to the minimum possible dwell time as this would result in minimal journey times if satisfiable. Therefore, the real task is the identification and resolving of conflicts resulting in a minimally relaxed yet valid timetable.

A conflict is an infeasible periodic event network. A conflict $\mathcal{C} = (\mathcal{V}, \mathcal{Z}, \vec{T})$ with $\mathcal{Z} \subseteq \mathcal{A}$ is called local conflict for $\mathcal{N}$ if and only if $\mathcal{C}$ is infeasible and $\mathcal{C}$ gets feasible by removing any constraint in $\mathcal{Z}$. A simple example conflict is outlined in figure 2. As the event network's constraints are encoded to separate clauses, local conflicts have a counterpart in SAT: A formula $\mathcal{M}$ in CNF is called minimally unsatisfiable subformula (MUS) if and only if $\mathcal{M}$ is unsatisfiable and $\mathcal{M}$ becomes satisfiable by removing any clause $c \in \mathcal{M}$. Likewise, there are highly efficient extractors for finding MUS [Ryv11]. Once a MUS is found, it can be decoded back to local conflicts [Gro12a].

Constraints in $\mathcal{Z}$ are relaxed by increasing the upper bound $t_{max,a}$ ($a \in \mathcal{Z}$), whereas symmetry constraints are relaxed by increasing maximum symmetry deviation d_a. Several

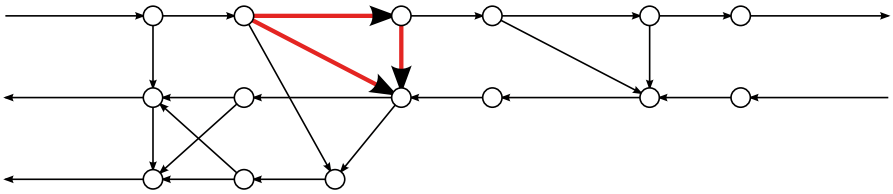


Figure 2: local conflict within a quite small sample event network

constraints, especially safety headway constraints, but also other constraints by user's request, are prohibited to be relaxed. In real world railway periodic event networks, the majority of constraints are unrelaxable headway constraints. Consequently, every conflict at first has to be checked on whether any relaxation of the relaxable arcs could solve the conflict. This is done by removing all relaxable arcs from the network and solving the remaining network. Remaining conflicts are intrinsic conflicts of the railway networks infrastructure and its operating program and thus, have to be manually resolved by modifying the operating program or the infrastructure. The extraction of local conflicts offers a detailed analysis of the bottlenecks [Opi09].

Resolving conflicts involves two steps: Firstly, the network is resolved by a quick heuristic, resulting in far too high relaxations. The most simple heuristic relaxes evenly all relaxable constraints by the same slack until the network is feasible. Afterwards, the relaxations are minimized under preservation of the network's feasibility. Minimization is done by either iterative usage of SAT solvers or direct usage of ILP solvers. The iterative process does not achieve the global minimum, but features much lower calculation times, whereas ILP solvers enable the use of more advanced objective functions.

Despite the impressive speedup achieved by using SAT solvers [Gro12b], a lot of timetabling problems are still too vast to be resolved directly in one piece in reasonable time. Therefore, strategies for an intelligent split of the timetabling problem is a necessity. In general, two methods were established: On the one hand, in hierarchical planning, the train network is sub-classified in several levels, for instance long-distance trains, local trains and freight trains as shown in figure 3. The trains of the highest level are scheduled first and then are left fixed, then the next level is scheduled and so on. This procedure represents the current manual timetabling processes well and easily fits with established paradigms. It reduces calculation time vastly, but it also cuts down the solution space remarkably.

On the other hand, a second method does not influence the solution space and also results in a considerable reduction of computation time. Some infeasible parts of the timetable network are extracted and resolved separately. Afterwards, all found out relaxations are adopted into the full network, which is then resolved again. Two automatic algorithms for determination of such network parts were developed: local conflict search, as described above and corridor analysis. Corridor analysis extracts the nodes and arcs of a route and its returning route and adds a defined amount of neighboring nodes and arcs. Far more sophisticated algorithms and

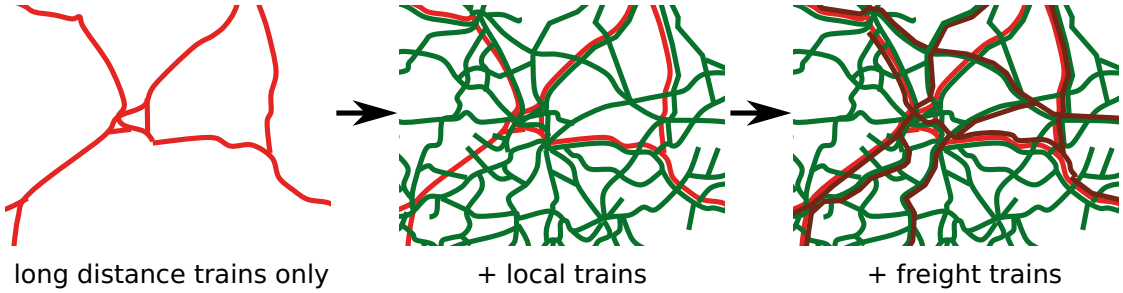


Figure 3: example for hierachical planning

combined strategies are currently under intensive research. Manual extraction of few heavily crowded and closely intertwined urban networks out of regional or national networks deliver effective results as well.

Both methods allow an intelligent and quick way of incremental scheduling, as they enable to fit small changes easily into existing timetables. Whereas solving whole networks takes hours, it is only a matter of minutes to include changes and to generate again valid timetables. Furthermore, it is possible to evaluate different operating programs and infrastructural states quickly.

5 Application and Results

As described in the beginning, the presented algorithms were implemented in the timetabling software system TAKT of the Chair of Traffic Flow Science at TU Dresden. The periodic event network is generated automatically from given input data. This is necessary, as large timetabling problems can consist of up to one million arcs and ten thousands of nodes, which cannot be calculated manually. The program automatically assigns the optimal route on the track layout to each train and calculates the running times within seconds. All minimum headways are calculated individually based on microscopical infrastructure data.

For instance, two passenger networks were studied: The German long-distance passenger railway network and the regional trains within the German region south-east (Saxony, Thuringia, Saxony-Anhalt). Firstly, the two networks were solved separately. As the network of regional trains is more dense, a particular complex part (Leipzig region) was extracted and the relaxtions needed for this part were calculated afore. Calculating a completely conflict-free timetable for the long-distance network takes about 2.5 hours. Determining a valid timetable for the regional trains in the described two iterations took approximately 2 hours each. In the joint network of both long-distance and regional trains, the long-distance trains were fixed to the determined timetable, whereas the regional trains were not fixed, but the relaxations computed before adopted to the event network. Solving the joint network took about 30 minutes. Providing fully correct input data, the program does not need any assistance to calculate applicable timetables. Having large amounts of input data, flaws like wrong

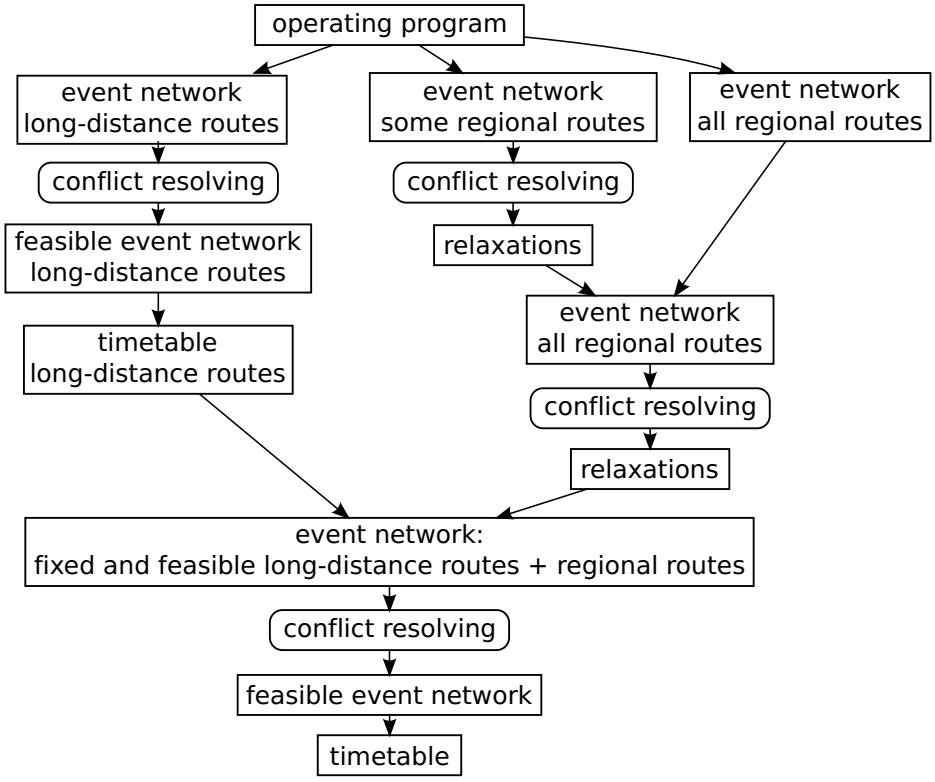


Figure 4: example for hierachical planning

train routes, missing stops, bogus connections are common. Thanks to the relaxation handling described before, the flaws can be detected and removed iteratively without calculating the timetable from scratch over and over again.

Based on PESP a more advanced model for completely automatic calculation of freight train paths along corridors is developed, that allows dynamic track allocations and dynamic selection of suitable speed profiles for freight trains. It outperforms the work of experienced experts even on highly crowded lines by far – same or even more train paths with a better quality are achieved in much lower time [Wei12; Opi09].

Additionally, TAKT features several visualization tools for evaluation of the resulting timetables, one example is displayed in figure 5.

6 Conclusion

As it was shown in this paper, SAT-based PESP solving and local conflict search provide a powerful base for fully automatic timetabling. Whereas prior approaches only solved small academic samples or needed significant simplification for solving real-world networks, the usage of SAT rose the size of calculable networks tremendously. Intelligent problem reduction and partition algorithms allow further increments in network size respectively reductions in calculation time. These improvements permit additional extensions to the PESP model as well. For instance dynamic track allocation for passenger trains will rise the flexibility of PESP

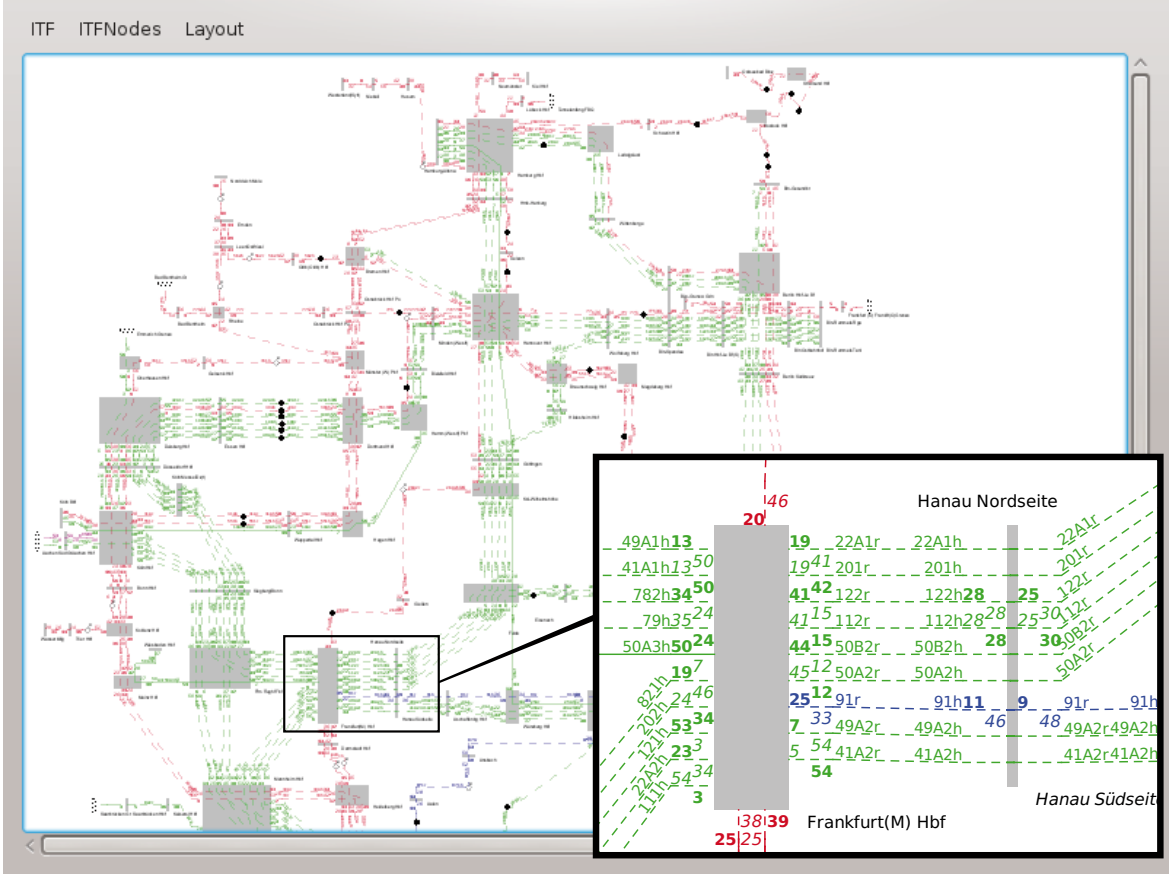


Figure 5: screenshot of timetable visualization in TAKT

in practical applications further.

The described management of relaxations offers an easy and efficient way for evolving timetables from scratch, which was successfully field-tested on large-scale railway networks. The possibility of fast rescheduling and the variety of realizable constraints opens periodic event scheduling for new fields like capacity research of several infrastructure states or railway traffic management in a completely new manner.

References

[Coo71] S. A. COOK: "The complexity of theorem-proving procedures". In: *Proceedings of the third annual ACM symposium on Theory of computing*. STOC '71. New York, NY, USA: ACM, 1971, pp. 151–158.

[Gro11] P. GROSSMANN: *Polynomial Reduction from PESP to SAT*. Tech. rep. 4. TU Dresden, Germany, Oct. 2011.

[Gro12a] P. GROSSMANN: *Extracting and Resolving Local Conflicts in Periodic Event Networks*. Diploma thesis. TU Dresden, 2012.

[Gro12b] P. GROSSMANN, S. HÖLLDOBLER, N. MANTHEY, K. NACHTIGALL, J. OPITZ, and P. STEINKE: "Solving Periodic Event Scheduling Problems with SAT". In: *Advanced Research in Applied Artificial Intelligence – 25th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems*. Vol. 7345. Lecture Notes in Computer Science. Springer, 2012, pp. 166–175.

[Lie06] C. LIEBCHEN: "Periodic Timetable Optimization in Public Transport". PhD thesis. TU Berlin. PhD thesis. Berlin, 2006. ISBN: 3-86624-150-X.

[Man10] N. MANTHEY and A. SAPTAWIJAYA: "Towards Improving the Resource Usage of SAT solvers". In: *POS-10*. Ed. by D. L. BERRE. Vol. 8. Edinburgh, 2010, pp. 28–40.

[Nac98] K. NACHTIGALL: *Periodic Network Optimization and Fixed Interval Timetable*. Habilitation thesis. Universität Hildesheim. 1998.

[Opi09] J. OPITZ: *Automatische Erzeugung und Optimierung von Taktfahrplänen in Schienenverkehrsnetzen*. PhD thesis. TU Dresden. Wiesbaden: Gabler Verlag, 2009. ISBN: 978-3-8349-2128-4.

[Ryv11] V. RYVCHIN and O. STRICHMAN: "Faster Extraction of High-Level Minimal Unsatisfiable Cores". In: *SAT*. Ed. by K. A. SAKALLAH and L. SIMON. Vol. 6695. Lecture Notes in Computer Science. Springer, 2011, pp. 174–187. ISBN: 978-3-642-21580-3.

[Ser89] P. SERAFINI and W. UKOVICH: "A Model for Periodic Scheduling Problems". In: *SIAM Journal on Discrete Mathematics* 2.4 (Nov. 1989), pp. 550–581.

[Tan11] T. TANJO, N. TAMURA, and M. BANBARA: "A Compact and Efficient SAT-Encoding of Finite Domain CSP". In: *SAT*. 2011, pp. 375–376.

[Wei12] R. WEISS, J. OPITZ, and K. NACHTIGALL: "A novel approach to strategic planning of rail freight transport". In: *OR Hannover, Presentation*. 2012.

Corresponding author: Michael Kümmling, Technische Universität Dresden, "Friedrich List" Faculty of Transportation and Traffic Sciences, 01062 Dresden, Germany, phone: +49 351 463 36506, e-mail: michael.kuemmling@tu-dresden.de

Analysis of a Closed-Loop Control Framework in a Realistic Railway Traffic Environment

Egidio Quaglietta, Francesco Corman, Rob M. P. Goverde
Delft University of Technology

Abstract

A wide literature is available on models and tools for the optimal real-time management of railway traffic, but the knowledge of their effects on real operations is still blurry and very limited due to the scarce implementation of these systems in practice. This paper analyses how these tools perform when interfaced in a closed-loop setup with a realistic traffic environment. A framework is developed that couples the rescheduling tool ROMA with the microscopic simulation model EGTRAIN. Railway traffic is managed for different perturbed scenarios using a rolling horizon scheme where optimal plans are periodically computed based on current traffic information and implemented in the simulation model. The closed-loop setup is investigated for different combinations of its parameters relatively to quality and stability of rescheduling plans. A comparison is performed against a typical open-loop approach that implements only the plan computed on the basis of expected train entrance delays. Both the closed-loop and the open-loop approaches are evaluated against the case in which no rescheduling is considered and trains keep on following the original timetable.

Results obtained for the Dutch corridor Utrecht-Den Bosch show that the closed-loop always outperforms the open-loop in terms of traffic performances. Short rescheduling intervals give more stable control strategies and higher quality improvements, but strongly increase computation times. Enlarging the prediction horizon beyond a given threshold do not improve the solution neither in terms of quality nor of stability.

Keywords: Real-time rescheduling, Closed-loop model predictive control, Stability analysis, Quality of dispatching plans.

1 Introduction

Railway operations are affected by unforeseen disturbances (e.g. extensions of dwell times at stations, unplanned stops at red signals) that induce deviations from the timetable and thereby reducing performances (e.g. punctuality). When time allowances in the timetable are

not enough to absorb such deviations it is necessary to reschedule railway traffic in real-time in order to mitigate the delay propagation and keep the capacity levels required by infrastructure managers. Railway dispatchers must therefore solve the so-called rescheduling problem, that is to find a plan (i.e. a combination of control measures like reordering, retiming and/or rerouting trains) that reduces the impact of delays on traffic. Such a plan is therefore called also a "solution" of the rescheduling problem.

In practice the rescheduling problem is currently solved on the basis of rules-of-thumb or the own experience of the dispatcher, with the aim of restoring the original timetable as soon as possible. These plans can be however ineffective or counterproductive due to the limited view that the human dispatcher has on downstream traffic behaviour. Advanced tools could be used instead that mathematically solve the rescheduling problem, providing to dispatchers plans that minimize the delay propagation on the network. In literature several models have been proposed so far for computing optimal rescheduling plans that guarantee operations free of track conflicts (where a conflict occurs when two trains want to occupy the same block section contemporarily). These approaches use different formulations for the rescheduling problem and adopt diverse objective functions and algorithms to solve it (see e.g. [Tör07], [Cor11], [Maz09]). The most of them are designed to be included within a rolling horizon setup (e.g. [Lüt09], [Cai12]) where at regular time intervals (*rescheduling interval RI*) current train information (e.g. measured speeds and positions) is used to predict track conflicts over a time period ahead (*prediction horizon PH*). If conflicts are detected a new conflict-free plan is computed.

Very few works (e.g. [Men11], [Tör07]) instead evaluate the quality of rescheduling solutions computed in a rolling horizon scheme considering the presence of stochastic traffic disturbances. However, the main shortcoming with such approaches is that no one has ever realized a closed-loop interaction (i.e. a bidirectional communication) between the rescheduling tool and a realistic traffic environment, to reliably evaluate the effects of optimal plans on train services. Practitioners are indeed still sceptic about using rescheduling tools into real operation, mainly because their implications on traffic are not investigated and not clear yet. This is also due to the scarcity of installations in practice (e.g. [Maz09], [Man09]) that prevent from having an extensive overview of their consequences.

This paper wants to clarify these issues by analysing the interaction of an optimal rescheduling tool with realistic traffic settings. We study a closed-loop rolling horizon setup for different configurations of the parameters RI and PH, evaluating the computed plans in terms of quality (i.e. effects on several measures of performance) and stability. A plan is defined as stable when it does not change if recomputed at later stages with respect to updated traffic information. A stable plan is therefore insensitive to the dynamic propagation of stochastic disturbances on the network. Stability is an essential requirement for rescheduling tools to prevent nervous behaviours of continuously changing solutions, that is hardly manageable by human dispatchers.

The effects of the closed-loop are then compared with those of a classic open-loop scheme in which the dispatcher only implements the plan computed at the beginning of the observation horizon on the basis of only the estimated train entrance delays. The benefits given by both the closed-loop and the open-loop rescheduling are assessed against the case

in which no rescheduling is applied at all and trains continue following the original timetable. The whole study is conducted over multiple disturbed scenarios and limited information on actual train dwell times.

A framework is developed that interfaces the state-of-the art rescheduling tool ROMA [Cor11] and the microscopic railway simulation model EGTRAIN [Qua11] , surrogate of the real field. The Dutch railway corridor Utrecht-Den Bosch is used as case-study.

In Section 2 the framework is described while the methodology is reported in Section 3. A practical application is reported in Section 4. Conclusions are supplied in Section 5.

2 Approach description

A closed-loop framework has been developed which connects the rescheduling tool ROMA (Railway Optimization by Means of Alternative Graphs) to a detailed stochastic microscopic model for the simulation of railway traffic, EGTRAIN (Environment for the desiGn and simulaTion of RAIlway Networks). EGTRAIN is considered realistic since it is validated by verifying that within undisturbed conditions simulated train running times were congruent with those scheduled in reality. Further research might include validation of the system for the full envelop of disturbed conditions. A detailed description of ROMA and EGTRAIN can be found respectively in [Cor11] and [Qua11].

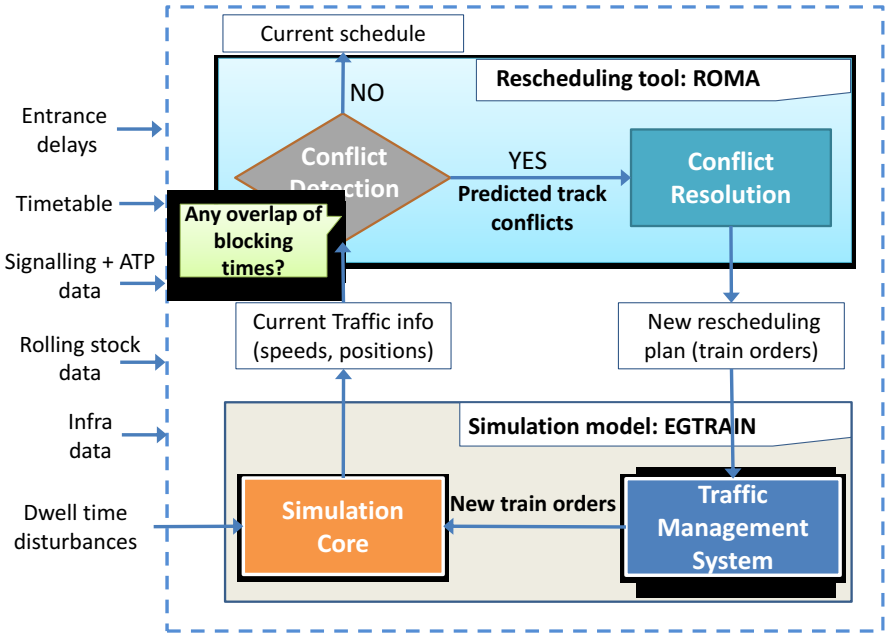


Figure 1. Architecture of the closed-loop framework.

As shown in Figure 1 both the rescheduling and the simulation models are initialized by specifying input data relative to the infrastructure, the rolling stock, the signaling and Automatic Train Protection (ATP) systems, the original timetable, and the entrance delays. To emulate a realistic traffic setting, random disturbances to dwell times are set only in the simulation model (since it represents the real field) but unknown to the rescheduling tool.

At a given time instant the simulation core of EGTRAIN sends current traffic information (positions and speeds of trains) to the Conflict Detection module of ROMA. Based on this information a deterministic prediction (i.e. train running and dwell times are considered as deterministic) of possible track conflicts is performed over a given period *PH.* Conflicts are identified by means of the blocking time theory [Han08] as overlaps between the blocking times of two trains for a certain block section. If no conflict is detected, the current schedule can still be operated without any modification. Otherwise, the predicted conflicts are sent as input to the Conflict Resolution module, which generates a new conflict-free plan by retiming (i.e. shifting the scheduled departure/arrival/passing times) and reordering (i.e. changing the passage order) trains in order to minimize the delay propagation on the network. This module represents the train scheduling problem as a job-shop model with no-store constraints that is solved by using a truncated version of a Branch and Bound algorithm [DAr07].

Train orders given by the new rescheduling plan at given locations (called checkpoint *CP*) are transferred to the Traffic Management System of EGTRAIN and implemented in the simulation core. Once implemented, the traffic is microscopically simulated (using a time-driven and synchronous approach) respecting the order supplied by the new plan for each specific location.

The interaction between the rescheduling and the simulation models follows a rolling horizon scheme (Figure 2). This means that the entire observation horizon *H* is subdivided in *n* successive stages, which are partially overlapping and spaced at regular time intervals called rescheduling intervals *RI*.

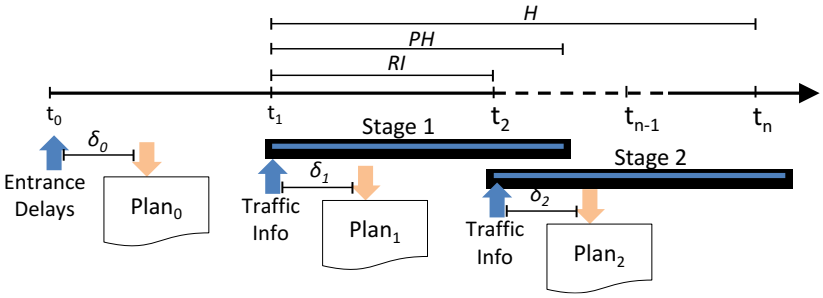


Figure 2. Rolling horizon scheme with inputs to ROMA (blue arrows) and to EGTRAIN (orange arrows) .

At the beginning of each stage (t_0, t_1,..., t_{n-1}) ROMA receives traffic information (considered not affected by measurements error) from EGTRAIN; predicts track conflicts over a prediction horizon *PH* that is constant for all stages, and provides (within the computing time δ_0, δ_1,..., δ_{n-1}) a new plan ($Plan_0$, $Plan_1$,..,$Plan_{n-1}$) that is implemented in EGTRAIN. In brief the complete closed-loop depicted in Figure 1 is performed after each RI. For the sake of simplicity we assume that the time to implement the plans is null, i.e. the simulation is frozen while ROMA computes, and the plans of ROMA are implemented in EGTRAIN as soon as they are computed.

The closed-loop setup has been tested for different combinations of *RI* and *PH* in order to understand how these parameters affect the performances of computed plans in terms of quality and stability.

A comparison is then performed against an open-loop approach that implements a rescheduling plan computed for the whole observation horizon *H*, only on the basis of the expected entrance delays. That is to say that the open-loop only puts into operation $Plan_0$ calculated by ROMA using a length of *PH* equal to the observation horizon (*PH*=*H*). In this case $Plan_0$ provides for the entire *H*, the solutions to all track conflicts that are expected to happen on the basis of only the entrance delays. This comparison consents us to evaluate which are the benefits given by the closed-loop when constantly updating the rescheduling plans with respect to current traffic conditions. In addition we also report what would happen if no rescheduling was applied at all, and trains operate according to the original timetable. In this way it is possible to understand which advantages the use of optimal rescheduling plans can bring to a situation in which no real-time management is considered.

The whole study is realized over different perturbed scenarios generated in a Monte-Carlo scheme, by randomly sampling: the entrance delays and disturbances to dwell times at stations. These latter are only considered in EGTRAIN and unknown to ROMA.

The metrics used for evaluating the stability of the rescheduling plans are:

Number of Relative Reordering (*NRR*). This metric describes for a certain location *CP* the similarity in terms of ordering between two plans computed at consecutive stages. Considering the plan given at stage *s*, we assume that a train is reordered if it is scheduled before some train that was preceding it, in the plan provided at stage *s*-1. The value of *NRR* is then calculated by counting all reordered trains.

The *average NRR* over all the rescheduling stages gives a measure of how stable in terms of reordering are the optimal plans provided by the rescheduling tool. The lower this average the higher is the plan stability. A condition of full stability is achieved when plans computed at consecutive stages are all the same, i.e. when the average NRR is zero.

The quality of all the plans (when traffic is rescheduled with the closed and the open loop) and the timetable (when no rescheduling is applied) is calculated with respect to the final station of trains by means of the following metrics:

Average total arrival delay (*AvTotDelay*). The total arrival delay of a train at a station is intended as the difference between the actual and the arrival time fixed by the original timetable at that station. *AvTotDelay* is the average of the total arrival delay over all delayed trains reaching their final station.

Average consecutive delay (AvConsDelay). For each train the consecutive delay at the final station is obtained by subtracting from its total arrival delay the unavoidable delays (i.e. entrance delays and dwell time disturbances cumulated at the previous stations). *AvConsDelay* is the average of this delay over all delayed trains reaching their final station. This metric gives a measure of how much trains are hindered during their run by the presence of other conflicting trains.

Max Consecutive Delay (MaxConsDelay) is the maximum value of the consecutive delay over all trains reaching their final station.

Punctuality at the final station with respect to a threshold of 3 (P_{3min})and 5 minutes (P_{5min}). These numbers give the percentage of trains whose total arrival delay at the final station is less than 3 and 5 minutes respectively.

3 Case Study: The Dutch corridor Utrecht-Den Bosch

The proposed framework is applied to the railway corridor between Utrecht (Ut) and Den Bosch (Ht) in the Netherlands. This has a length of more than 48 km with 6 intermediate stations: Lunetten (Ln), Houten (Htn), Houten Castellum (Htnc), Culemborg (Cl), Geldermalsen (Gdm), and Zaltbommel (Zbm). The schematic layout is presented in Figure 3, together with the locations in which trains can overtake each other and a reordering is possible (CP1, CP2, CP3). The network is equipped with a fixed-block signalling system and the traditional Dutch automatic train protection ATB system. The hourly periodic timetable schedules 4 intercity trains (IC) per hour per direction between Ut and Ht without intermediate stops; and 4 regional trains, two of which are limited between Ut and Gdm, while the other two run all the way till Ht. No freight trains are taken into account in the study. For the sake of simplicity, only trains running along the Ut-Ht direction are considered, as in this double-track corridor there is no interaction between trains running in opposite directions. The observation horizon in which the rescheduling is applied is H= 120 min. The closed-loop setup has been tested for 9 different parameter combinations obtained by coupling 3 values of *RI*: 30, 60 and 120 s, with 3 lengths of the *PH*: 15, 30 and 60 min. The only solution ($Plan_0$) implemented within the open-loop has been calculated by adopting a *PH* equal to the whole observation horizon, i.e. PH = 120 min. The study is performed over 30 different perturbed scenarios obtained by sampling: *i*) entrance delays from a Weibull distribution fitted to real data [Cor11] with scale, shape and shift parameters that are different for ICs and regional trains; *ii*) station dwell times have been considered normally distribution with a lower truncation to the minimum dwell time, the planned duration as mean, and 60% of this latter as standard deviation; this distribution results in a cumulative delay over all stops that is averagely 1.5 to 2 min per train, in accordance to reality.

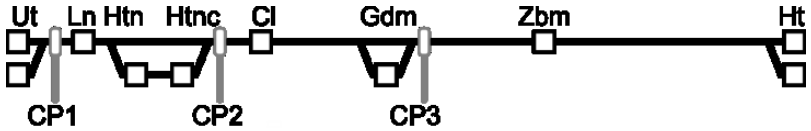


Figure 3. Schematic layout of the Utrecht - Den Bosch corridor, with the locations (CP1, CP2 and CP3) in which train reordering is considered.

3.1 Results

The results obtained for all the stability and quality metrics are computed as the average over the 30 disturbed scenarios. Figure 4 shows how the rescheduling plans vary over time in terms of *NRR* for different *RI*s and *PH*s of the closed-loop setup. For a given stage the value of *NRR* is aggregated over the three CPs, i.e. it is the sum of their corresponding *NRR*. For the first 18 minutes the rescheduling solution is practically stable and equal to $Plan_0$, i.e. the plan computed on the basis of only expected entrance delays. This is because in this period only two trains have entered the network and stochastic disturbances have not propagated yet. As such disturbances start progressing over the network, the rescheduling plans become unstable and vary over time. The reason of such instability is that the propagation of disturbances induces a deviation between actual and predicted train trajectories, altering from time to time the conflicts detected by ROMA and the corresponding solutions (i.e. the

plans). For a fixed *RI*, the variation in terms of train reordering *NRR* is higher for longer *PH*s. For example when fixing *RI*=30 (see Figure 4*a*), this average has a very strong increase of 109% when extending the *PH* from 15 to 30 min and then only a slight increment of 11% when further enlarging the *PH* to 60 min. The same behaviour is shown for the other tested values of *RI* (see Figure 4*b*-4*c*). These results suggest that for a fixed *RI* the plan stability decreases when enlarging the *PH*, until a threshold τ (in this case τ =30 min) beyond which it remains more or less constant. The motivation is that shorter *PH*s are less affected by prediction errors since only the closest future is estimated. Moreover in this case only a limited knowledge is available of traffic evolution and time margins exploitable for reordering. In this myopic situation the rescheduling tool can mostly solve conflicts by retiming (i.e. propagating delays to later trains) rather than reordering, as verified in [Qua13].

This explains why the value of *NRR* at a certain stage is generally lower for shorter *PH*s. For longer *PH*s, conflict predictions are more uncertain (therefore more variable), given that more errors are possible when estimating traffic over a farther future. When progressively enlarging the *PH* it will be achieved a threshold length τ beyond which computed plans do not consistently differ since traffic predictions (and their errors) are basically the same.

Although the presence of sharper peaks in the value of *NRR*, more stable plans (hence more easily manageable by human dispatchers) are obtained for short *RI*s. In this case the average *NRR* is indeed lower than the one relative to larger *RI*s. This is because smaller errors affect the prediction if this latter is updated more frequently on the basis of current train information. For example for *PH*=30 min, such average increments of 30% when enlarging *RI* from 30 to 60 s. When *RI* is widened from 60 to 120 s, a smaller increase of 19% is instead observed.

Table 1. Quality indices for the different traffic management approaches.

RI [s]	PH [min]	AvTot Delay [s]	AvCons Delay [s]	MaxCons Delay [s]	P_{3min} [%]	P_{5min} [%]	$T_{EGTRAIN}$ [s]	T_{ROMA} [s]
Timetable	n/a	118.30	28.31	107.42	87.45	89.52	56.37	n/a
Open-loop	120	106.57	23.52	105.84	90.15	92.79	56.66	1.82
120	15	102.08	22.02	96.64	90.33	93.13	57.50	91.97
	30	100.72	21.41	95.44	90.61	93.24	57.31	92.87
	60	100.72	21.41	95.44	90.61	93.24	58.36	98.37
60	15	99.85	18.02	76.20	90.33	93.96	57.30	178.36
	30	97.51	16.58	71.20	90.78	94.27	57.86	188.07
	60	97.51	16.58	71.20	90.78	94.27	57.91	196.94
30	15	94.36	15.21	77.47	90.61	94.85	57.10	322.75
	30	94.24	15.07	68.65	90.91	94.85	57.57	333.66
	60	91.65	14.52	68.65	91.19	94.85	56.81	357.44

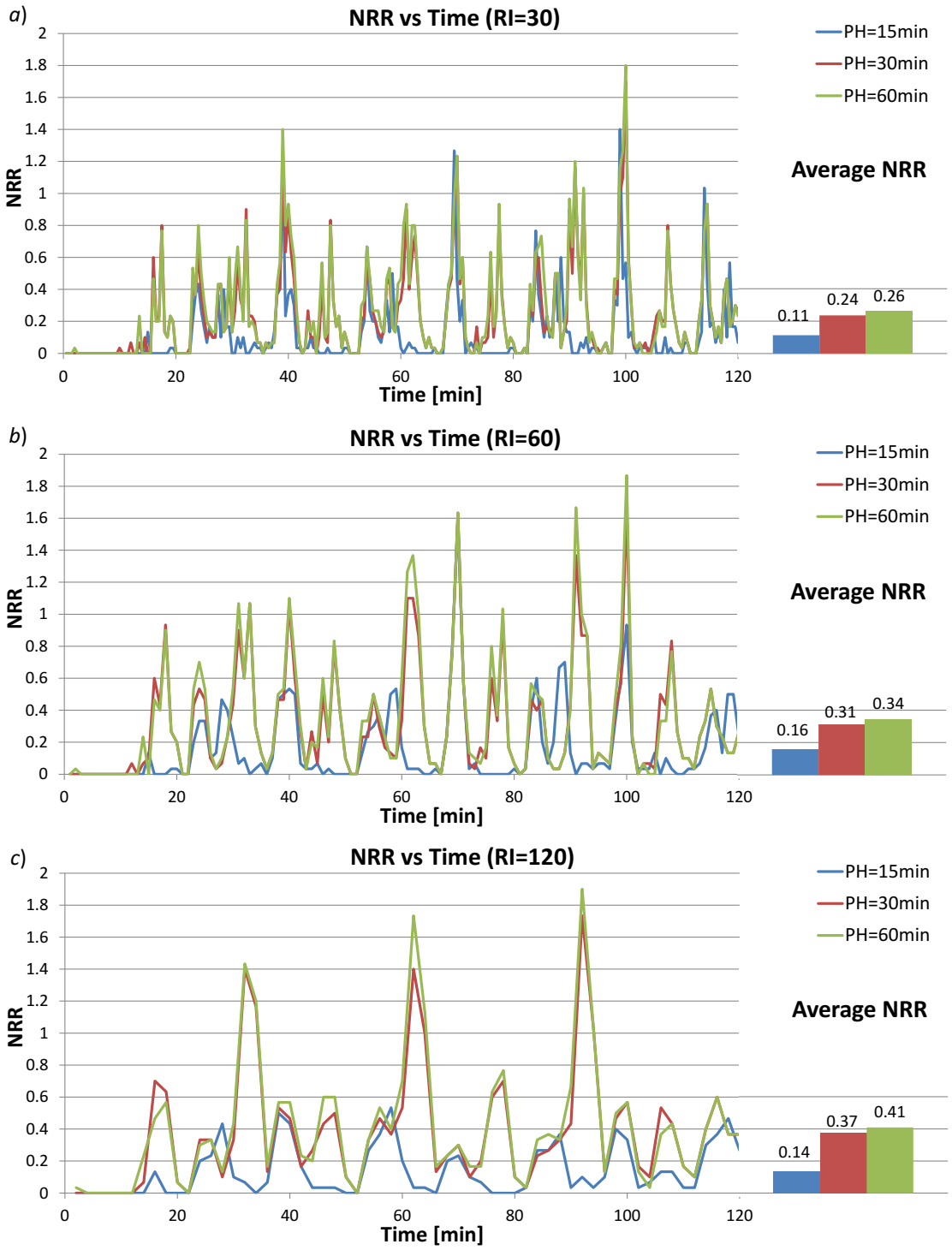


Figure 4. NRR and average NRR (aggregated for all the CPs) for the different configurations of the closed-loop setup.

In Table 1 the effects on traffic are reported in terms of the mentioned quality indices for the timetable, the open-loop and the different configurations of the closed-loop. The last two columns report the total computation time for simulation (by EGTRAIN) and for rescheduling (by ROMA); this latter is in average 1.5 second per stage.

This table clearly highlights the benefits of implementing optimal rescheduling plans

instead of leaving traffic operating according to the timetable. A large improvement in traffic performances is already reached when adopting the open-loop approach. In this case we obtain a reduction of *AvTotDelay*, *AvConsDelay* and *MaxConsDelay* that is respectively of 10%, 17% and 1.5% with respect to the timetable. Consistent gains are also achieved in punctuality since the number of punctual trains increases of 21.5% for the threshold of 3 min and 31.2% for the one of 5 min. Larger improvements are achieved when applying the closed-loop rescheduling. Indeed the closed-loop outperforms the open-loop for all tested combinations of its parameters *RI* and *PH.* For instance the closed-loop with *RI*=120 s and PH=15 min improves the open-loop solution of 4.5%, 6.4%, 8.7% respectively for the three measures of delay while 2% and 4.8% in terms of punctual trains at 3 and 5 min. When the *PH* is enlarged to 30 min these measures of performance are further improved respectively of: 1.3%, 3%, 1.2%, 2.9% and 1.6%. Widening the *PH* up to 60 min no improvement is instead observed. For a fixed value of *RI*, we can say that the quality of rescheduling solutions improves when enlarging the *PH* until the threshold value of 30 min. Beyond this value the improvement seems to be null (as in the case of *RI*=120 and 60s) or only marginal (when *RI*=30s). Very short *PH*s (i.e. 15 min) are less effective than larger ones since the rescheduling tool is forced to solve conflicts mainly by retiming rather than reordering. On the other hand, *PH*s larger than the threshold of 30 min can only marginally improve the solution, while certainly increasing the total computation time of the rescheduling tool (reported in the column T_{ROMA} in Table 1). This conclusion is fully in line with what previously deduced by Törnquist in [Tör07].

The improvement of the solution is much more sensitive to the variation of *RI* than to the one of *PH*. When fixing for example the *PH* to 30 min, the closed-loop with *RI* = 120 s improves the open-loop solution of 6%, 9%, 10%, 4.7% and 6.3%, respectively for *AvTotDelay*, *AvConsDelay, MaxConsDelay,* and the amount of punctual trains at 3 and 5 min. When *RI* is reduced to 60 s, such measures of performance are further improved respectively of: 3.2%, 22.5%, 25.4%, 2% and 15%. If *RI* is further reduced to 30 s, these performances are still improved of 3.4%, 9.1%, 8.8% 1.4% and 10%. The closed-loop setup with short *RI* heavily improves the quality of the rescheduling plans with respect to an open-loop approach. In this case the critical point is constituted by the total computation time of the rescheduling tool that practically doubles each time that *RI* is reduced. The total simulation time $T_{EGTRAIN}$ is instead more or less constant and averagely equal to 57.34 s. The value of *RI* that guarantees the best performances of the closed-loop setup must be chosen on the basis of an optimal trade-off between solution quality and total computation time.

4 Conclusions

This paper presents an innovative analysis of a closed-loop rolling horizon approach for the optimal real-time management of railway traffic. A framework has been developed that dynamically integrates the tool for optimal rescheduling ROMA, with the microscopic railway traffic simulation model EGTRAIN, that is considered as a valid substitute of the real field. A practical application is realized to the Dutch railway corridor Utrecht-Den Bosch.

Results underline the beneficial impacts on traffic that optimal rescheduling can bring with respect to the case in which no rescheduling is applied and trains keep on following the original timetable. The closed-loop rescheduling approach always outperforms the open-loop. Specifically we observed that the solution quality strongly improves when shortening the *RI* of the closed-loop, although the computation times of the rescheduling tool heavily increase. The choice of the best value for *RI* must therefore allow a satisfactory trade-off between solution quality and computation times. A smaller role has instead the *PH* which improves solution quality if not too short. On the other hand *PH*s longer than a threshold τ bring only marginal improvements while increasing computation times. As for quality, the closed-loop shows a similar behaviour for the stability of its plans. Indeed short *RI*s give on average more stable plans in terms of train reordering, although they vary more sharply. Short *PH*s return slighter variations in the plans since in this case less reordering is performed. Plan stability is more or less constant while enlarging the *PH*s over a threshold τ.

The main conclusion of this study on closed-loop setups is the recommendation for a short value of *RI* and a length of the *PH* beyond which the quality of the plans do not consistently improve anymore. Preliminary studies are advised to identify for each specific case these values of *RI* and *PH*.

Future research will be addressed to determine these values for different case-studies and how the closed-loop performs in the case of both heavy and slight perturbations. Moreover we will investigate the impacts on traffic performances when plans of the closed-loop are implemented after a certain time needed by the dispatcher to practically communicate them to the field.

Acknowledgments

The research contained in this paper is partly supported by the EU FP7 project ON-TIME (www.ontime-project.eu).

References

[Cai12] G. Caimi, M. Fuchsberger, M. Laumanns, and M. Lüthi: “A Model Predictive Control Approach For Discrete-Time Rescheduling In Complex Central Railway Station Areas”. In: *Computers & Operations Research* 39 (2012), pp. 2578-2593.

[DAr07] A. D’Ariano, D. Pacciarelli, and M. Pranzo: “A Branch And Bound Algorithm For Scheduling Trains In A Railway Network”. In: *EJOR* 183.2 (2007), pp. 643–657.

[Cor11] F. Corman, A. D’Ariano, M. Pranzo and I. A. Hansen: “Effectiveness of dynamic reordering and rerouting of trains in a complicated and densely occupied station area”. In: *Transportation Planning and Technology* 34.4 (2011), pp. 341-362.

[Han08] I. A. Hansen and J. Pachl: *Railway Timetable and Traffic*. Hamburg: Eurailpress, 2008.

[Lüt09] M. Lüthi: “Improving the efficiency of heavily used railway networks through

integrated real-time rescheduling". PhD thesis. Zürich: ETH Zurich, 2009.

[Man09] C. MANNINO and A. MASCIS: "Optimal Real-Time Control in Metro Stations". In: *Operations Research* 57.4 (2009), pp. 1026-1039.

[Maz09] M. MAZZARELLO and E. OTTAVIANI: "A Traffic Management System For Real-Time Traffic Optimisation In Railways". In: *Transportation Research Part B* 41.2 (2007), pp. 246-274.

[Men11] L. MENG and X. ZHOU: "Robust Single-Track Train Dispatching Model Under A Dynamic And Stochastic Environment: A Scenario-Based Rolling Horizon Solution Approach". In: *Transportation Research Part B* 45 (2011), pp. 1080-1102.

[Qua11] E. QUAGLIETTA: "A Microscopic Simulation Model For Supporting The Design Of Railway Systems: Development And Applications". PhD thesis. Naples: University of Naples Federico II, 2011.

[Qua13] E. QUAGLIETTA, F. CORMAN, and R. M. P. GOVERDE: "Impact of a stochastic and dynamic setting on the stability of railway dispatching solutions". In: *16th International IEEE conference on ITS*, Den Haag, the Netherlands, Oct. 6-9, 2013.

[Tör07] J. TÖRNQUIST: "Railway Traffic Disturbance Management-An Experimental Analysis Of Disturbance Complexity, Management Objectives And Limitations In Planning Horizon". In: *Transportation Research Part A* 41 (2007), pp. 249-266.

Corresponding author: Egidio Quaglietta, Delft University of Technology, Department of Transport and Planning, 2628 CN Delft, The Netherlands, phone: +31 15 2784914, e-mail: e.quaglietta@tudelft.nl

Boosting the Performance of a MILP Formulation for Railway Traffic Management in Complex Junctions

Paola Pellegrini, Grégory Marlière, Joaquín Rodriguez
Ifsttar – ESTAS, Université Lille Nord de France

Abstract

Unexpected events often perturb railway traffic. The impact of these events may be very remarkable in terms of delay propagation. We analyze a mixed integer linear programming (MILP) formulation which aims at minimizing the delay propagation when traffic is perturbed. It does so by modifying train routing and scheduling at junctions. This formulation is able to solve to optimality many realistic instances in a computation time which is in line with real-time purposes. However, for the most difficult instances, finding the optimal solution is too time consuming. In this paper, we assess the performance of the MILP formulation when a short time limit is imposed. Moreover, we propose different methods for boosting this performance. We tackle instances representing traffic in the Lille-Flandres station (France), and we show that the boosted MILP formulation achieves very positive results, finding the optimal solution in more than 75% of the experimental runs.

Keywords: real-time railway traffic management problem, mixed-integer linear programming, track-circuit, complex junction

1 Introduction

Railway traffic is often perturbed by unexpected events which cause primary delays. These primary delays may cause the emergence of conflicts: when a train does not respect its original schedule, it may claim a track section in concurrence with another train; one of these trains must then slow down, or even stop, to ensure safety. Hence, one of these trains will suffer a secondary delay due to the traffic perturbation: secondary delay is the delay that trains incur into due to the emergence of conflicts, opposed to the primary delay that is due to unexpected events, such as the presence of snow on the tracks.

The emergence of conflicts is particularly remarkable at junctions, that is, at locations where multiple lines cross. Here, several routes are often available for connecting an entry to

an exit line: train routes may possibly be changed with respect to the originally chosen ones, still respecting the train origins and destinations. Furthermore, the fact that different lines cross allows interventions on the train schedule: dispatchers may decide to impose a precise train order at critical locations. The secondary delay deriving from a traffic perturbation may potentially be strongly limited through wise routing and scheduling decisions.

In the practice, these decisions are mostly made manually by the dispatchers. However, a noticeable number of academic studies have recently been devoted to finding effective algorithms for real-time railway traffic management (see, e.g., [Cai12; Cor09; Lus12; Maz07; Pel13; Rod07; Tor07]).

In previous works [Pel12; Pel13], we proposed a mixed-integer linear programming (MILP) formulation for solving to optimality the problem of routing and scheduling trains in case of railway traffic perturbation, using a fixed-speed model and representing the route-lock sectional-release interlocking system. Although very often this MILP formulation quickly finds the optimal solution to realistic instances, it fails sometime in delivering it within a computation time in line with real-time purposes.

In this paper, we study the performance of the MILP-based heuristic. We obtain it by running the MILP formulation proposed in our previous works for a limited computation time: we quit the search process after this time has elapsed. We assess the ability of the MILP-based heuristic to find the optimal solution within the time limit, and its error rate when it fails to do so. Moreover, we propose different methods for boosting the MILP-based heuristic performance, and we show that the results achieved on instances representing traffic on the Lille-Flandres station (France) become extremely promising.

The rest of the paper is organized as follows. In Section 2, we present the MILP formulation. In Section 3, we present the methods proposed for boosting its performance. In Section 4 and 4, we report the experimental setup and the results of the analysis, respectively. Finally, in Section 6, we draw conclusions.

2 Mixed-integer linear programming formulation

In the MILP formulation, we model the infrastructure in terms of track-circuits, that is, into track sections on which the presence of a train is automatically detected. In addition to the existing track-circuits, we introduce two dummy ones: tc_0 and tc_∞. They represent the entry and the exit locations of the infrastructure, respectively. Sequences of track-circuits are grouped into block sections, which are opened by a signal indicating their availability. Before a train can enter (start the occupation of) a block section, all the track-circuits belonging to the same block section must be reserved for the train itself. In the following, we will name the sum of reservation and occupation time as utilization time. If a train starts its trip at null speed from a platform, then we consider the beginning of the occupation to correspond to the moment in which the train starts moving. If it remains still at the platform, its actual utilization will be ensured in the model through reservation. We consider the case of the signal opening the block section having three possible aspects (green, yellow and red). In the

model, this translates into the need for the train to reserve two consecutive block sections before being allowed to enter the first of them. Moreover, each block section is reserved by the train some time before its entering, to allow the route formation, and it remains reserved after its leaving, to allow the route release. Finally, we consider routes which do not include any stop within their starting and ending point. If a train is scheduled to stop at an intermediate point of the infrastructure, we double this train into two trains using the same rolling stock. In the MILP formulation, we use this notation:

T, R, TC, PL	set of trains, routes, track-circuits and platforms ($PL \subset TC$),
ty_t	type corresponding to train t (indicating train characteristics),
$e(tc, r)$	indicator function: 1 if track-circuit tc belongs to an extreme (either the first or the last) block section on route r, 0 otherwise,
R_t, TC_t	set of routes and track-circuits which can be used by train t,
TC^r	set of track-circuits composing route r,
$bs_{r,tc}$	block section including track-circuit tc along route r,
$p_{r,tc}$, $s_{r,tc}$	track-circuits preceding and following tc along route r,
$ref_{r,tc}$	reference track-circuit for the reservation of tc along route r,
$TC(tc, tc', r)$	set of track-circuits between tc and tc' along route r,
$rt_{ty,r,tc}$, $ct_{ty,r,tc}$	running and clearing time of tc along r for a train of type ty,
for, rel	formation and release time,
$init_t$, $sched_t$	earliest time at which train t can be operated, and earliest time at which train t can reach its destination given $init_t$ and the route assigned to t in the timetable,
$i(t, t')$	indicator function: 1 if trains t and t' use the same rolling stock and t' results from the turnaround, join or split of train t, 0 otherwise,
ms	minimum separation between the arrival of a train and the departure of another train which uses the same rolling stock,
M	large constant.

We define **continuous variables**, all non-negative:
for all trains $t \in T$:

$$D_t = \text{secondary delay suffered by train } t;$$

for all triplets of train $t \in T$, route $r \in R_t$ and track-circuit $tc \in TC^r$:

$$o_{t,r,tc} = \text{time in which } t \text{ starts the occupation of } tc \text{ along route } r,$$

$$d_{t,r,tc} = \text{delay suffered by } t \text{ in } tc \text{ along route } r \text{ (defined if } bs_{r,tc} \neq bs_{r,s_{tc,r}});$$

for all pairs of train $t \in T$ and track-circuit $tc \in TC_t$:

$$sU_{t,tc}, eU_{t,tc} = \text{time in which } tc \text{ starts and ends being utilized by } t, \text{ respectively.}$$

Moreover, we define **binary variables**:

for all pairs of train $t \in T$ and route $r \in R_t$:

$$x_{t,r} = 1 \text{ if } t \text{ uses } r; 0 \text{ otherwise},$$

for all triplets of train $t, t' \in T$ and track-circuit $tc \in TC_t \cap TC_{t'}$:

$$y_{t,t',tc} = 1 \text{ if } t \text{ utilizes } tc \text{ before } t' \ (t \prec t'); 0 \text{ otherwise } (t \succ t').$$

The objective function which we consider in this paper is the minimization of the total secondary delay suffered by trains:

$$\min \sum_{t \in T} D_t. \tag{1}$$

Indeed, the objective function could capture different train priorities, which may play a role in practical railway traffic management: the secondary delay suffered by each train may be multiplied by a factor representing its importance. The total delay is to be minimized while respecting the following sets of constraints:

$$o_{t,r,tc} \geq init_t \, x_{t,r} \quad \forall t \in T, r \in R_t, tc \in TC^r : \tag{2}$$

trains cannot be operated earlier than $init_t$;

$$o_{t,r,tc} \leq M x_{t,r} \quad \forall t \in T, r \in R_t, tc \in TC^r : \tag{3}$$

the start of track-circuit occupation along a route is zero if the route itself is not used;

$$o_{t,r,tc} \geq o_{t,r,p_{r,tc}} + rt_{r,ty_t,p_{r,tc}} x_{t,r} \quad \forall t \in T, r \in R_t, tc \in TC^r : \tag{4}$$

a train cannot start occupying track-circuit tc along a route if it has not spent in the preceding track-circuit at least its running time, if the route is used;

$$\sum_{r \in R_t} x_{t,r} = 1 \quad \forall t \in T : \tag{5}$$

exactly one route is used by each train;

$$D_t \geq \sum_{r \in R_t} o_{t,r,tc_\infty} - sched_t \quad \forall t \in T : \tag{6}$$

variables D_t must be coherent with the actual and the scheduled arrival times;

$$d_{t,r,tc} = o_{t,r,s_{r,tc}} - o_{t,r,tc} - rt_{r,ty_t,tc} x_{t,r} \quad \forall t \in T, r \in R_t, \ tc \in TC^r : \ bs_{r,tc} \neq bs_{r,s_{r,tc}} : \tag{7}$$

the delay variable is equal to the time in which t starts occupying the track-circuit following

tc, minus the time in which it starts occupying tc itself, minus the running time;

$$\sum_{\substack{r\in R_t, tc\in TC^r:\\ p_{r,tc}=tc_0}} o_{t,r,tc} \geq \sum_{\substack{r\in R_{t'}, tc\in TC^r:\\ s_{r,tc}=tc_\infty}} o_{t',r,tc} + (ms + rt_{r,ty_{t'},tc})x_{t',r} \quad \forall t,t' \in T : i(t',t) = 1 : \quad (8)$$

a time ms must separate the arrival and departure of trains using the same rolling stock. If two trains are in connection, an equivalent constraint is to be imposed;

$$\sum_{r\in R_t, tc\in TC^r: p_{r,tc}=tc_0} sU_{t,tc} \leq \sum_{r\in R_{t'}, tc\in TC^r: s_{r,tc}=tc_\infty} eU_{t',tc} \quad \forall t,t' \in T : i(t',t) = 1 : \quad (9)$$

the track-circuit where the turnaround, join or split takes place is reserved by t' until it arrives at the platform, plus the release time, and then it is immediately reserved by t;

$$\sum_{r\in R_t: tc\in TC^r} x_{t,r} = \sum_{r\in R_{t'}: tc\in TC^r} x_{t',r} \quad \forall t,t' \in T : i(t',t) = 1, tc \in PL : \quad (10)$$

trains using the same rolling stock must use routes including the same platform;

$$sU_{t,tc} = \sum_{r\in R_t: tc\in TC^r} \left(o_{t,r,ref_{r,tc}} - for\, x_{t,r}\right) \quad \forall t \in T,\ tc \in TC_t : (\nexists\, t' \in T : i(t',t) = 1) \vee (\forall\, r \in R_t :\ ref_{r,tc} \neq s_{r,tc_0}) : \quad (11)$$

a train's utilization of a track-circuit starts as soon as the train starts occupying the track-circuit $ref_{r,tc}$ along one of the routes including it, minus the formation time. If we are considering a track-circuit of the first two block sections of the route ($ref_{r,tc} = s_{r,tc_0}$) and the concerned train t results from the turnaround, join or split of one or more other trains, we must impose these constraints as inequalities. This is due to the need of keeping platforms utilized. In fact, if t results from the turnaround of t', Constraint (9) ensures that the platform where the turnaround takes place is starts being reserved by t as soon as t' arrives. However, t needs to wait at least for a time ms before departing. The occupation of the platform by t is however considered starting form its actual departure, for guaranteeing the coherence of the occupation variables and the running time (Constraints (4)). Hence, t's reservation starts much earlier than its occupation;

$$eU_{t,tc} = \sum_{r\in R_t: tc\in TC^r} o_{t,r,ref_{r,tc}} + ul_{t,r,tc} \quad \forall t \in T,\ tc \in TC_t : \quad (12)$$

the utilization of a track-circuit tc lasts as long as the train utilizes it along any route ($ul_{t,r,tc}$, which includes the running time of all track-circuits between $ref_{r,tc}$ and tc, the delay and the release time), plus the formation time;

$$y_{t,t',tc} + y_{t',t,tc} = 1 \quad \forall t,t' \in T,\ tc \in TC_t \cap TC_{t'}, \quad (13)$$

$$eU_{t,tc} - M(1 - y_{t,t',tc}) \leq sU_{t',tc} \quad \forall t,t' \in T : tc \in TC_t \cap TC_{t'} : \quad (14)$$

$$eU_{t',tc} - My_{t,t',tc} \leq sU_{t,tc} \quad \begin{array}{c} i(t,t')\sum_{r\in R_t} e(tc,r) = 0 \ \wedge \ i(t',t)\sum_{r\in R_{t'}} e(tc,r) = 0, \\ \forall t,t' \in T : tc \in TC_t \cap TC_{t'} : \\ i(t,t')\sum_{r\in R_t} e(tc,r) = 0 \ \wedge \ i(t',t)\sum_{r\in R_{t'}} e(tc,r) = 0 : \end{array} \quad (15)$$

track-circuit utilizations by two trains do not overlap. These constrains are disjunctive ones, as those used, for example, by Törnquist and Persson [Tor07]. For further discussion on the formulation, we refer the reader to our previous works [Pel12; Pel13].

The main difference between this formulation and the ones that have been proposed in the literature, as the one by Corman et al. [Cor09], is that here we can model the route-lock sectional-release interlocking system. Instead, Corman et al. [Cor09] report a model based on alternative graphs considering either the route-lock route-release or the sectional-lock sectional-release interlocking system. The authors state that a sophistication of the model can be used to consider the route-lock sectional-release interlocking system, but they do not present the details of such a sophistication.

3 Performance boosting methods

In this paper, we propose different methods for boosting the performance of the MILP formulation within a fix time limit: the MILP-based heuristic.

The first boosting method consists in changing the setting of one parameter of the solver that we use in the computational analysis: IBM ILOG CPLEX Concert Technology for C++ (IBM ILOG CPLEX version 12) [IBM12]. In particular, we set the **MIP emphasis switch parameter** (MIPEmphasis) to 4. It controls the trade-offs between speed, feasibility, optimality, and moving bounds. By default, CPLEX works toward a rapid proof of an optimal solution, but balances that with effort toward finding high quality feasible solutions early in the optimization. Here, we impose that CPLEX works hard to find high quality feasible solutions. When using this parameter setting, we will add "m" to the algorithm reference.

The three other boosting methods are algorithmic expedients. First of all, we apply a **backward shift** to all time references: we move both $init_t$'s and $sched_t$'s backwards of a time interval equal to $\min_{t\in T} init_t - for$. Second, we include the solution process in a **two optimization step cycle**. In the first step, we perform an optimization imposing the use of the routes fixed in the timetable. In the second step, we use the solution so obtained as starting point for the optimization with all possible train routes. Third, we exploit the solution obtained in the first optimization step for **reducing the value of *M***, the large constant used in the formulation. For ensuring the coherence of Constraints (14) and (15), the value of the constant M needs to be at least equal to the latest end of a concerned track-circuit utilization. In the first optimization step, we set M conservatively high, for ensuring this coherence. Let S^{*1} be the optimal solution in the first optimization step. Let $\sum_{t\in T} D_t^{*1}$ be the total delay associated to S^{*1}. When increasing the number of available routes, all solutions improving over S^{*1} will have, by definition, an associated total delay smaller than $\sum_{t\in T} D_t^{*1}$. Hence, even if the whole delay was assigned to a single train, its latest reservation of track-circuit tc

would be at most equal to the sum of: the maximum across all available routes of the earliest possible exit time from tc along each route; the difference between the length of the shortest and scheduled route (which might not be the shortest one, but with respect to which $sched_t$ is computed); $\sum_{t \in T} D_t^{*1}$. In a constraint involving more than one train, we set M equal to the maximum of these quantities computed for the trains involved on the concerned track-circuit. When applying these expedients, we will add "s", "2" and "M" to the algorithm reference, respectively.

4 Experimental setup

As mentioned in Section 3, we implementes the MILP formulation and the boosting methods through the IBM ILOG CPLEX Concert Technology for C++. We ran the experiments on Intel Xeon 2.67GHz processor with 24 GB RAM, under Linux Ubuntu distribution version 12.04., and we executed CPLEX excluding parallel computation. For each run, we imposed a limit of either 1 or 3 minutes of CPU time, which is in line with real-time purposes [Rod07].

We tackle instances representing perturbations of the timetable of a weekday in 2002 in the control area including the main station of Lille in the North of France, i.e., the Lille-Flandres station. In particular, we consider a Wednesday timetable including 589 trains. Being the Lille-Flandres a terminal station, all rolling stocks are used for both an arriving and a departing train, but for what concerns the first trains departing in the morning (which arrived the day before to the platform) and the last ones arriving at night (which will leave the platform the day after). Besides 259 turnarounds, the timetable contains 8 joins and 10 splits. The station is linked to seven regional, national and international lines and it has 17 platforms. A total of 2409 routes exist and they are composed by 299 track-circuits. We consider formation and release times of 15 and 5 seconds, respectively. Starting from the original timetable, we impose a delay to 20% of trains that do not represent shunting movements: we randomly select the trains to be delayed and we randomly draw their delay in the interval between 5 and 15 minutes [Lus12]. Both these random selections are based on uniform probability distributions. By replicating the random assignment of train primary delay 30 times, we obtain 30 different perturbed one-day timetables. For each of these 30 perturbed one-day timetables, we solve ten 60-minute instances, randomly drawing the starting time of ten time horizons between 5 am and 5 pm: we tackle 300 instances including 25 to 50 trains (mean 31). In these experiments, we consider instances independently from one another, neglecting the transition between consecutive time horizons. For a discussion of how such a transition can be implemented, we refer the reader to Pellegrini et al. [Pel12].

For each instance, we perform 15 runs for each version of the algorithm considered: we test all the combinations of the boosting methods discussed in Section 3, considering that if a single optimization step is performed, then the reduction of M makes no sense.

m → MIP emphasis switch parameter =4
s → backward shift
2 → two optimization step cycle
M → reduction of the value of M

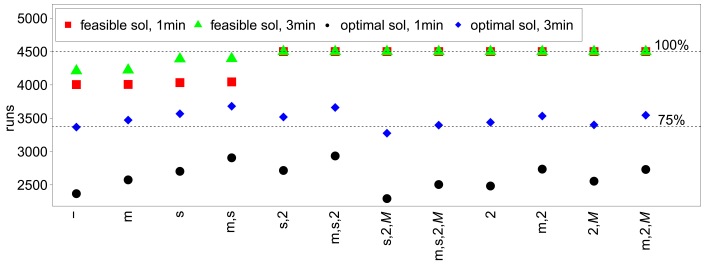


Figure 1: Number of runs in which each setup finds a feasible and an optimal solution.

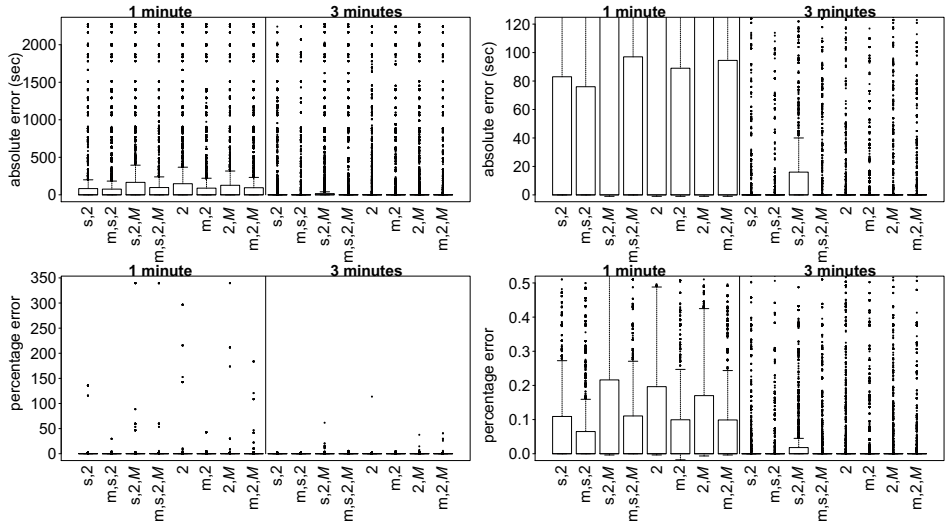


Figure 2: Absolute (top) and percentage (bottom) error. Whole distribution (left) and zoom on the y-axis (right).

5 Experimental results

In this analysis, we assess the performance of the algorithm obtained by running the MILP formulation described in Section 2 for a short computation time: the MILP-based heuristic. Moreover, we assess the performance improvements achieved when applying the boosting methods proposed in Section 3.

As mentioned in Section 4, we tackle 300 instances and we perform 15 runs for each of them. Hence, we employ each version of the MILP-based heuristic 4500 times. Figure 1 reports the number of instances in which a feasible or an optimal solution is found within the time limit of either 1 or 3 minutes. Only eight versions always find at least one feasible solution: all the versions excluding the two optimization step cycle sometime fail in delivering any solution. This is a major failure, and hence we do not consider them in the rest of the analysis. As for the eight successful versions, the figure shows that the number of runs in which an optimal solution is found increases over time, as expected, and it attains a very high level (always around 75%) within 3 minute computation.

Figure 2 depicts through boxplots the distributions of the absolute (value of the best

Table 1: Statistical significance according to the Wilcoxon rank-sum test with confidence level 0.95. A 1 (3) indicates a significant difference in favor of the setup indexing the line with respect to the one indexing the column, in runs of 1 (3) minute(s).

	s,2	m,s,2	s,2,M	m,s,2,M	2	m,2	2,M	m,2,M
s,2	-		1 3	1 3	1 3		1 3	3
m,s,2	1 3	-	1 3	1 3	1 3	1 3	1 3	1 3
s,2,M			-					
m,s,2,M			1 3	-	1		1	
2			1 3		-			
m,2	3		1 3	1 3	1 3	-	1 3	3
2,M			1 3		1		-	
m,2,M			1 3	1 3	1 3		1 3	-

solution found minus value of the optimal solution) and percentage (value of the best solution found minus value of the optimal solution, divided by the latter) error made by the successful versions of the MILP-based heuristic within 1 and 3 minutes, with respect to the optimal solution. Both from an absolute and a percentage perspective, sometimes rather high errors are registered, but the error is null in the great majority of the cases: the solution returned in 1 minute computation is often the optimal one, and this is true even more often in 3 minute computation. In many cases, the boxplots do not allow the identification of differences between the various versions. However, these differences exist, and Table 1 reports the statistical significance of the difference between each couple of setups. The meaning of this significance, for example in the case of m,s,2 being significantly better than s,2, is that, if we consider further instances with similar characteristics to the ones used here, we can expect to achieve a better solution through m,s,2 than through s,2. As the second line of the table shows, the setup including the appropriate setting of the MIP emphasis switch parameter, the backward shift and the two optimization step cycle (m,s,2) is always better than the other setups in statistical terms. Even if it common knowledge that a small value of M contributes to the strength of a model, in these experiments we do not detect any advantage brought by the reduction of the value of M: the versions of the MILP-based heuristic including this boosting method are often outperformed by the versions excluding it. The backward shift appears useful, but actually not crucial: what really makes the difference is the two optimization step cycle. Finally, despite being less influential than this latter boosting method, the appropriate setting of the MIP emphasis switch parameter positively contributes to the performance improvement of the MILP-based heuristic.

6 Conclusions

In this paper, we have proposed different methods for boosting the performance of a MILP-based heuristic for the problem of routing and scheduling trains in case of railway traffic perturbation. The MILP-based heuristic consists in solving instances through a MILP formulation, interrupting the search process after a fix time limit.

We performed a thorough experimental analysis based on instances representing one hour traffic at the Lille-Flandres station, in France, to assess the impact of the boosting methods

individually and in combination with each other within either 1 or 3 minute computation. The results show that the performance of the MILP-based heuristic are very promising when the appropriate boosting methods are implemented.

In future research, we will test the impact on the performance of the introduction of valid inequalities in the model. Moreover, we will more deeply focus on the parameter settings of the MILP solver for identifying possible further performance improvements achievable thanks to a more appropriate configuration.

References

[Cai12] G. CAIMI, M. FUCHSBERGER, M. LAUMANNS, and M. LÜTHI: "A model predictive control approach for descrete-time rescheduling in complex central railway station approach". In: *Computers & Operations Research* 39 (2012), pp. 2578–2593.

[Cor09] F. CORMAN, R. GOVERDE, and A. D'ARIANO: "Rescheduling dense traffic over complex station interlocking areas". In: *Robust and Online Large-Scale Optimization: Models and Techniques for Transportation Systems*. Ed. by R. A. et AL. Vol. 5868. LNCS. Berlin, Germany: Springer, 2009, pp. 369–386.

[IBM12] IBM CORPORATION: *User's manual for CPLEX*. 2012.

[Lus12] R. LUSBY, J. LARSEN, M. EHRGOTT, and D. RYAN: "A set packing inspired method for real-time junction train routing". In: *Computers & Operations Research* 40(3) (2012), pp. 713–724.

[Maz07] M. MAZZARELLO and E. OTTAVIANI: "A traffic management system for real-time traffic optimisation in railways". In: *Transportation Research Part B* 41 (2007), pp. 246–274.

[Pel12] P. PELLEGRINI, G. MARLIÈRE, and J. RODRIGUEZ: "Real Time Railway Traffic Management Modeling Track-Circuits". In: *ATMOS 2012*. Ed. by D. DELLING and L. LIBERTI. Vol. 25. OASIcs. Dagstuhl, Germany: Leibniz-Zentrum fuer Informatik, 2012, pp. 23–34.

[Pel13] P. PELLEGRINI, G. MARLIÈRE, and J. RODRIGUEZ: "A mixed-integer linear program for the real-time railway traffic management problem modeling track-circuits". In: *5th ISROR conference, RailCopenhagen 2013*. Copenhagen, Denmark, 2013.

[Rod07] J. RODRIGUEZ: "A constraint programming model for real-time train scheduling at junctions". In: *Transportation Research Part B* 41 (2007), pp. 231–245.

[Tor07] J. TÖRNQUIST and J. PERSSON: "N-tracked railway traffic re-scheduling during disturbances". In: *Transportation Research Part B* 41 (2007), pp. 342–362.

Corresponding author: Paola Pellegrini, Ifsttar – ESTAS, rue Élisée Reclus 20, 59666 Villeneuve d'Ascq, France, phone: +33 3 20 43 84 04, e-mail: paola.pellegrini@ifsttar.fr

Railway Traffic Control with Minimization of Passengers' Discomfort

Francesco Corman[1], Federico Sabene[2], Dario Pacciarelli[2], Marcella Samà[2], Andrea D'Ariano[2]

[1] Delft University of Technology and Katholieke Universiteit Leuven

[2] Università degli Studi Roma Tre

Abstract

In the recent literature on real-time train rescheduling, at least two main research lines can be distinguished. On the one hand, rescheduling approaches focus on the feasibility of disposition schedules for practical applications. Thus, the models in this research line tend to incorporate as many practical details as it is necessary to ensure the schedule feasibility in practice, while the objective function is typically the minimization of train delays. Some of these approaches are currently being implemented in practice. On the other hand, delay management approaches focus more on the customer point of view, and tend to manage the rail service in real time in order to minimize the discomfort of the passengers. Models in this research line tend to be simplified, while the main focus is on the design of the objective function, which is typically related to the minimization of passengers' delays. This work combines the two approaches by incorporating the passengers' point of view into a detailed train rescheduling model. The overall problem is decomposed into two optimization problems, namely the rescheduling of trains by taking into account the number of passengers per train and the rerouting of passengers by taking into account the train schedule. An illustrative example shows the approach. Computational experiments on a preliminary test case, using a commercial solver, show that the approach is very promising to increase railway customer satisfaction.

Keywords: train rescheduling, real-time railway traffic control, delay management, transit assignment, MILP.

1 Introduction and literature review

Railway service is a key factor to reduce congestion on highways and other means of transport, especially in densely populated areas, and to provide an eco-friendly and sustainable way of transport. In order to attract new customers from other transport modes it is particularly important to improve the quality of service (QoS) offered to the passengers.

Despite the many efforts aiming at improving the QoS through carefully designed offline plans (timetable), delays and other sources of passengers' discomfort are experienced every day. The originating causes of disruptions and perturbations, such as bad weather conditions or power outages, cannot be always avoided and result in primary delays. Real-time rescheduling aims at mitigation of the consequences of primary delays, i.e., at minimization of the secondary delays caused by widespread propagation of primary delays. Thus, any small improvement in the performance of real-time rescheduling has a direct positive impact on the QoS perceived by passengers. This fact motivates the remarkable amount of research recently to the development of advanced decision support systems for real-time railway traffic management.

The complexity of the train rescheduling problem stems from the limited overtaking capacity of railway lines and the constraints of the safety system, such as the signal status and speed restrictions. One of the most effective approaches to tackle such complexity is based on the blocking time theory and on the alternative graph model [Mas02]. Advanced scheduling approaches based on these concepts are able to quickly solve real-life train instances in which train arrival times, orders and routes, are considered variable (see e.g. [Dar07, Man09, Cor10, Cor11]). There are, however, other promising approaches based on MILP formulations [Tor11]. All these approaches focus on the practical feasibility of the schedules produced and are able to manage train traffic in practical size networks within a computation time compatible with real-time operations. The objective functions typically focus on train delays and the solutions produced demonstrated remarkable improvement with respect to the current practice and/or to the basic dispatching rules adopted in most practical applications. One weakness of all these models is the limited view of passenger needs and expectations, which are taken into account only indirectly, i.e., by penalizing train delays, platform changes or broken transfer connections. Among the works trying to enlarge the scope of these approaches, [Cor11] proposes an iterated lexicographic optimization of train delays, given a division of trains into classes. The delay of each class is minimized provided that the delay of higher priority classes does not increase. This approach might be applied by defining priority classes according to the estimated importance of particular trains for the overall passenger QoS. A biobjective optimization approach is proposed by [Cor12], in which a weight is associated to each passenger connections, depending on its importance for the passenger QoS, and then the Pareto frontier is computed where the two objective functions are the train delays and the total weight of broken connections.

One stream of research which directly faces the optimization of the QoS perceived by the passengers is based on the concept of customer-oriented dispatching [Suhl01]. Among this stream of research, the delay management problem [Sch07, Dol12] decides whether to keep or not transfer connections during operations, a crucial decision for passenger flows. The approaches in this stream of research are currently based on macroscopic models, whose major drawback is the gap between the QoS promised by the optimal solutions and the one achieved when implementing the solutions in practice.

A combination of the delay management approach with the microscopic models based on the alternative graph concept is proposed by [Cor13], in which passengers delays are optimized by iteratively solving a microscopic scheduling problem (without knowledge of

passenger flows) and a macroscopic delay management problem (without explicit modelling of limited infrastructure capacity).

We address the problem of finding microscopically feasible train schedules minimizing passenger delays. Differently from [Cor13], scheduling and traffic control decisions are taken by a single model. Compared to [Tam13], a detailed microscopic optimization model for rescheduling is used together with a comprehensive assignment of passengers to time-space paths in the network. The solution procedure, at each iteration, alternates a rescheduling phase, in which train orders and times are optimized for given assignment of passengers to train services, to a passengers rerouting phase, in which passengers' delay is minimized for the given train schedules. In this phase passengers are assigned to the shortest-path in a time-distance graph for each origin-destination pair. The procedure alternates the two phases until convergence (i.e., until the optimal solution in one phase is equivalent to the solution found in the previous iteration).

A formal description of the problem is provided in Section2; the proposed model is explained with the help of an example in Section 3. Section 4 introduces a test case and evaluation of the models; Section 5 gives conclusions and future research directions.

2 Problem definition

Train traffic is typically planned through a detailed timetable, defined months in advance, which satisfies the expected passenger demand by suitable choice of lines, arriving/departure times and transfers between train services at major stations. During their services, trains run in the network following their given routes and are supposed to obey to published departure times. However, delays and disturbances often occur that cause a positive difference between the realized and published arrival time, thus requiring rescheduling decisions to be taken. In this context, primary delays are caused by external disturbances that can be recovered only at a certain extent by exploiting running time supplements. Secondary delays are determined by rescheduling decisions in response to primary delays, and are necessary to solve train conflicts.

To ensure safe movements and no collision between trains, the railway network is divided into block sections, where only one train at a time is allowed. Signalling and safety systems regulate train movements by allowing access to at most one train at a time to each block section. A conflict occurs whenever two trains require the same block section at the same time. The train rescheduling problem consists in solving all conflicts by finding a passing order for trains at each block section and platform of the network in a given time horizon.

We make the following simplified assumptions. Trains have infinite capacity and each passenger chooses the route allowing him/her to reach his/her destination as soon as possible, called in the following the shortest OD path. We assume to know the number of passengers willing to reach the same destination D from the same origin O in the same time window W (e.g., the time between two consecutive departures from the same origin). With these assumptions, all passengers in a triple (O,D,W) will follow the same shortest OD path (we assume this path as unique, possibly breaking ties arbitrarily).

2.1 Optimization model for scheduling railway traffic

Most optimization models for train rescheduling associate trains to jobs and block sections to machines. Thus, the problem corresponds to a job shop scheduling problem with blocking no-swap constraints [Mas02]. The latter constraint ensures that a train on a given block section cannot move forward if the block section ahead is not available (e.g. if occupied by another train). This prevents any other train to enter the given block section. More details on job shop based models for railway traffic can be found on [Cor11].

The entrance of a train in a block section is an operation, and its minimum starting time h is a decision variable. Operations are associated to the traversing of block sections by each train, as well as to the dwell time in each station. For each operation is given a minimum processing time p, equal to the traversing time of the associated block section or to the minimum dwell time in a station. Hence, the starting times of successive operations i and j of the same train are linked by a *fixed* constraint $t_j \geq t_i + p_i$. Let F be the set of fixed constraints. Conflicting operations of different trains on the same block section need to be separated by a minimum headway separation s. A passing order x for the two trains must also be chosen. If k, i are successive operations of a train, j, h are successive operations of another train and k and j are two conflicting operations associated to the entrance of the two trains in the same block section, then there is a pair of *alternative* constraints $(h_j \geq h_i + s_{ij})\text{OR}(h_k \geq h_h + s_{hk})$. The first inequality of the alternative pair corresponds to the precedence given to k over i, since j can start only s_{ij} time units after the completion of k (i.e., after the start of i). The second inequality corresponds to i preceding k. The MILP formulation of a pair requires to introduce a binary variable x_{ijhk} associated to the choice of the precedence between the conflicting operations, equal to 1 if k precedes j and 0 otherwise. P is the set of alternative pairs. As for the objective function to be used for the scheduling problem, we propose the minimization of the total weighted tardiness (z) of the trains at a set E of due-date points, each point e associated with a planned arrival time p_e, that include the published stops and the exit from the network. The weight f_e corresponds to the expected number of passengers on the train at point e. The resulting rescheduling model reads as follows:

$$
\begin{aligned}
&\min \sum_{e \in E} f_e z_e \\
&\begin{cases}
h_j \geq h_i + p_{ij} & (i,j) \in F \\
h_j \geq h_i + s_{ij} - M\left(1 - x_{ijhk}\right) & \\
h_k \geq h_h + s_{hk} - M x_{ijhk} & ((i,j),(h,k)) \in P \\
z_e \geq h_e - p_e & e \in E \\
h, z \geq 0; \quad x \in \{0,1\}^{|P|} &
\end{cases}
\end{aligned}
$$

2.2 Transit assignment model

The transit assignment problem forecasts the distribution of passengers onto the railway network. The simplified behavioural assumptions made on the infinite capacity of the trains and on the passengers' reaction to disturbances make possible to decompose the transit assignment problem and to solve it independently for each triple (O,D,W). Given a timing h

for the trains at each station, it is possible to build a time-space graph $G=(N,A)$ in which nodes in N correspond to the arrival/departure time of a train at/from some station, i.e., each node $i \in N$ is associated to operation i defined in Section 2.1, with associated time h_i. There are two types of arcs in G: there is a *travel* arc in G for each train service between two consecutive stopping stations. Arc (i,j) is therefore associated to a departure time h_i and an arrival time h_j. Besides travel arcs, there are *waiting* arcs $(h_i h_j)$ in G between the arrival time h_i of a train at some station and the departure time h_j of some train (the same or another) from the same station, with $h_j > h_i$. Each arc has a weight equal to $(h_j - h_i)$. With this definition, all passengers in a triple (O,D,W) will choose the shortest path in G from a node OW, associated to the departure time $h_{OW} \in W$ of the first train leaving O, to a node DW, associated to the arrival time h_{DW} of the first feasible train arriving in D. Note that G is acyclic and is the same for all (O,D,W) pairs. Letting n_{ODW} be the number of passengers of a triple (O,D,W), and q_{ij}^{ODW} a flow variable equal to 1 if arc (i,j) belongs to the shortest OD path and 0 otherwise, the transit assignment problem for a triple (O,D,W) can be formulated as a MinCostFlow problem:

$$
\begin{aligned}
&\min \sum_{(i,j)\in A} n_{ODW}\left(h_j - h_i\right) q_{ij}^{ODW} \\
&\begin{cases}
\sum_{(i,j)\in \delta^{out}(OW)} q_{ij}^{ODW} = 1 \\
\sum_{(i,j)\in \delta^{in}(DW)} q_{ij}^{ODW} = 1 \\
\sum_{(i,j)\in \delta^{out}(i)} q_{ij}^{ODW} = \sum_{(i,j)\in \delta^{in}(i)} q_{ij}^{ODW} \qquad i \in N \setminus \{OW, DW\} \\
q_{ij}^{ODW} \in \{0,1\} \qquad (i,j) \in A
\end{cases}
\end{aligned}
$$

2.3 Integration and illustrative example

The passenger assignment model is inspired by [Dol12]; the main difference is that here times of operations (h) are not decision variables, but rather the result of the optimization performed in a stage of the scheduling model. This simplifies the transit assignment problem and allows for faster computation, even if in principle there is no guarantee that the final solution produces the global minimum passengers' discomfort.

The overall algorithm alternates the solution of a rescheduling problem and a transit assignment problem. We start by solving the train rescheduling problem with $f_e = 1$ for all $e \in E$. Based on the rescheduling solution, we perform the assignment of passengers to trains by choosing the shortest OD path for each triple (O,D,W) in the updated traffic situation. The expected passenger flows are then used to update the value f_e in the rescheduling problem. The new solution to the rescheduling problem is then again used to compute a new assignment, and so on iteratively until convergence. For the small test cases here considered, convergence is always achieved within the third iteration.

We next illustrate the application of the algorithm on an illustrative network, based on the fictional network shown in Figure 1. We consider a single (O,D,W) triple that originates in the leftmost station and has its destination in the rightmost station, as shown by the Origin and Destination reported in Figure 1. Passengers appears at the Origin with a constant rate of 10

passengers per minute. The timetable considers 5 trains (labelled A-E), with different stopping patterns (identified by a box in correspondence of a station) and two operating speeds (trains A and C are fast, in solid lines in Figure 1; Trains B D and E are slow and depicted with dotted lines); trains of different categories can overtake each other at each station if there are platforms available for overtaking.

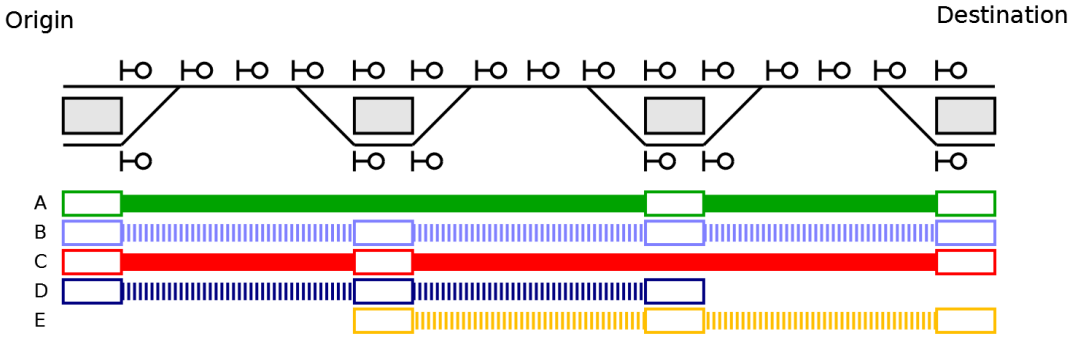


Figure 1: Network and stopping pattern considered.

Figure 2 (left) reports the time-distance plot of the solution found at the first iteration of the rescheduling problem: time is on the y axis, increasing upwards; distance on the x-axis. Though the infrastructure available allows for overtaking at stations, this possibility is not exploited in this schedule. The average total train delay is 9 minutes, while the average consecutive delay is 2 minutes. This solution is optimal concerning the maximum consecutive delay, that equals 8 minutes.

Given this schedule, the solution of the transit assignment problem is given in Figure 2(right) in terms of the resulting flow of passengers. At time 60, passengers in the Path OD_4 leave train D to board on train C in order to reach their destination. The total travel time of the 580 passengers is 50100 minutes. As a comparison, note that the fastest train take 47 minutes to go from origin to destination, therefore a lower bound to the optimum passengers' travel time is 27260 minutes.

At the second step, a new instance of the train rescheduling problem is solved with the updated values f_e of Figure 2 (right). The new rescheduling solution is shown in Figure 3 (left). Passengers on train D board train C in order to reach the destination in the fastest way possible. The average total delay increases to 13 minutes; the average consecutive delay to 7minutes. This increase is mainly due to Train E, which is scheduled last, since in this solution there are no passengers on this train. This is of course a consequence of the simplified ODW considered in this illustrative example, while reality shows more complex patterns of passenger demand. Train A overtakes train D at the second station, as the former transport more passengers than the latter.

Figure 3(right) reports the new optimal assignment of passengers to train services associated to the new schedule. The resulting total travel time for the 580 passengers is now 47900 min, 5% less than the first iteration. At the 3rd rescheduling iteration, the same schedule of iteration 2 is obtained, convergence is thus reached and the procedure is completed.

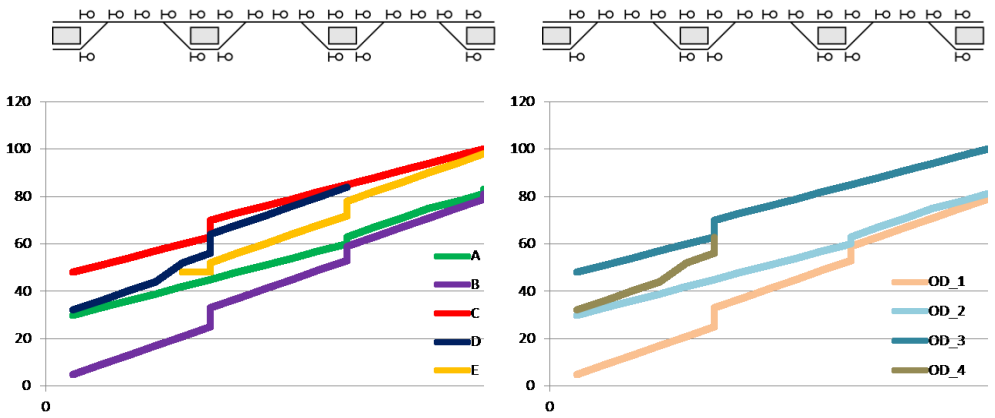


Figure 2: Time-distance plot for 1st iteration of train rescheduling (left); Corresponding transit assignment solution (right).

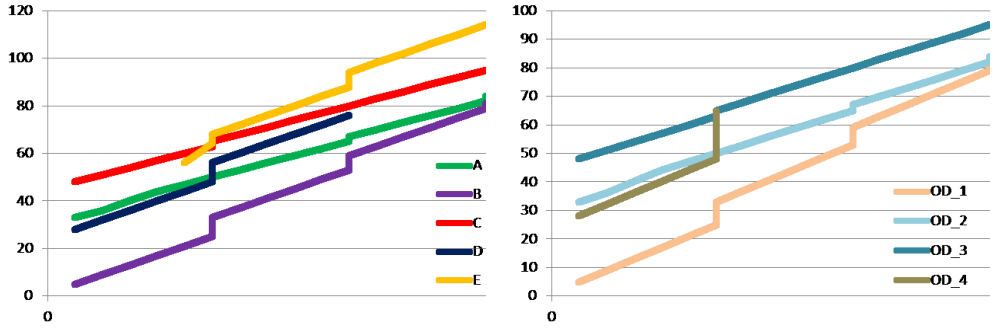


Figure 3: Time-distance plot for 2nd iteration of train scheduling (left); Corresponding transit assignment solution (right).

3 Experimental assessment

Based on the infrastructure and timetable given in Figure 1, we next report on some preliminary computational results, based on 5 random delay instances (every train is subjected to a uniform random delay between 0 and 15 minutes). We evaluate three different approaches. The first is the First-In-First-Out (FIFO) dispatching rule, that assigns a shared block section to the first train requiring it. The FIFO rule is a common benchmark for the assessment of train rescheduling algorithms. The second approach is the solution minimizing the maximum consecutive delay, adopted by the ROMA system (see e.g. [Cor12]). The third one is the approach proposed in this paper (labelled PaxFlows), for which the results achieved at the first two iterations are provided. For this toy example, convergence is always achieved within two iterations. The computation time to solve at optimum with CPLEX the scheduling and assignment problem was always well below 1 second.

In Table 1, we report the averages over the 5 instances, describing the solutions in terms of a variety of performance indicators, namely: the total travel time of passengers, $\sum_{(i,j)\in A} n_{ODW}\left(h_j - h_i\right)q_{ij}^{ODW}$, the weighted train delay $\sum_{e\in E} f_e z_e$, the average total delay per train;

the average consecutive delay per train, and the maximum consecutive delay, all in minutes. From the results we can conclude that, even for this small example, the minimization of the total travel time of passengers allows a sharp reduction of passenger discomfort: passenger travel time is reduced by 10% compared to ROMA and by 12% compared to FIFO; a strong reduction is also achieved with regard to the weighted train delay. As expected, some performance indicators show a degree of conflict, since decreasing the passenger travel time may result in an increase of the maximum consecutive delay. This can also be due to the fact that in this example there are no running time supplements for the trains.

Table 1: Performance of the approaches regarding passenger and train delays.

Model	Passengers Travel Time [min]	Weighted train delay [min]	Average total delay [min]	Average consecutive delay [min]	Maximum consecutive delay [min]
FIFO	45128	5103	9.7	2.2	8.4
ROMA	44190	5988	11.3	3.6	5
PaxFlows 1st iter	40598	12348	8.9	1.3	5
PaxFlows, 2nd iter	40014	4413	10.4	2.9	12.8

Analysing the iterative approach PaxFlows, the passenger travel time decreases slightly across the iterations even in this very simple test case; the average total delay suffers from an increase of 16%; the average and maximum consecutive delay are both more than doubled. This is due to the fact that trains carrying few passengers never get the precedence over trains carrying many passengers. Concerning the convergence of the algorithm, for the considered instances two iterations are always sufficient to reach convergence, and the second iteration only provides minor adjustments to the objective function, while the first transit assignment step provides the most significant benefits. The improvement in terms of passenger travel time is already 8% at the first iteration, compared to ROMA that minimizes exclusively delay propagation.

The results suggest that it is worth considering passenger flows in the rescheduling phase in order to reduce the passengers' discomfort. Optimization approaches with a global view address better passengers' discomfort than local myopic rules such as FIFO. We also found that rescheduling models can consider more elaborated objective functions without losing too much in terms of computation times.

4 Conclusions and further work

This paper moves a step forward in the integration of rescheduling and delay management techniques. An iterative approach is proposed, solving a train rescheduling problem and a passenger assignment problem in sequence, until convergence is reached. Experiments on a small fictional test case demonstrate the potential of this approach for increasing the passenger satisfaction and the quality of railway service.

The next steps would require to carry out experiments with realistic test cases of suitable size and complexity, including a larger set of OD pairs. A larger set of real-life delays would allow a more extensive evaluation of the approach proposed, as well as the possibility to generalize results to other networks.

Different research directions should focus on the increase of computational efficiency to allow for real-time usage, and on the design of suitable exact or heuristic algorithms for solving the decomposed problem defined in this paper. It would be also interesting to analyse the exact solution of the problem as a whole, i.e. without decomposition into rescheduling and passenger assignment. In principle, the integration of the two models into a single MILP model is easy, but it would be interesting to see how this would influence the computation time of solution algorithms and the quality of the resulting solutions.

More sophisticated assignment models could be studied, e.g., by considering the finite capacity of each train, more accurate measures of the passengers' discomfort, or other approaches in the literature [Mes07]. For example, the time spent by each passenger on the train or in the station could be weighted differently, the passenger discomfort while traveling on a train could be increasing with the number of passengers traveling on the same train, or even could be taken into consideration the anxiety of the passengers as a factor of discomfort, which increases with the train delay as the risk of missing a connection with the next train increases.

A further set of interesting open challenges concerns the design of more realistic models for the passengers' behaviour: how are passengers going to react when a change to their preferred path is suggested? Are they going to stick to their (offline) decision, are they going to follow the suggestion for alternative modes of connectivity (i.e. another line), or disregard travelling at all? Approaches based on discrete choice theory might be helpful to model these questions, and analysis of recorded passenger flows might provide insights on how passenger flows react to unexpected events [Hur12]. This might increase the degree of realism of the passenger assignment and forecast of the OD pairs.

Acknowledgements

We acknowledge support from by the State Key Laboratory of Rail Traffic Control & Safety (contract No RCS2012K004), Beijing Jiaotong University; and by COST Action TransITS TU1004: Modelling Public Transport Passenger Flows in the Era of ITS.

References

[An08] J. AN, J. TENG, and L. MENG: "A BRT network route design model". In: *IEEE Conference on Intelligent Transportation Systems, Proceedings.* Beijing, 2008.

[Cor10] F. CORMAN, A. D'ARIANO, D. PACCIARELLI, and M. PRANZO: "A tabu search algorithm for rerouting trains during rail operations". In: *Transportation Research Part B* 44.1 (2010), pp. 175–192.

[Cor11] F. CORMAN, A. D'ARIANO, HANSEN, and PACCIARELLI: "Optimal multi-class rescheduling of railway traffic". In: *Journal of Rail Transport Planning & Management* 1.1 (2011), pp. 14–24.

[Cor12] F. CORMAN, A. D'ARIANO, D. PACCIARELLI, and M. PRANZO: "Bi-objective conflict detection and resolution in railway traffic management". In: *Transportation Research Part C: Emerging Technologies* 20.1 (2012), pp. 79–94.

[Cor13] F. CORMAN, T. DOLLEVOET, A. D'ARIANO, and D. HUISMAN: "An iterative optimization framework for delay management and train scheduling". In: *XXVI EURO-INFORMS conference.* Rome, July 2013.

[Dar07] A. D'ARIANO, D. PACCIARELLI, and M. PRANZO: "A branch and bound algorithm for scheduling trains in a railway network". In: *European Journal of Operational Research* 183.2 (2007), pp. 643–657.

[Dol12] T. DOLLEVOET, D. HUISMAN, M. SCHMIDT, and A. SCHÖBEL: "Delay management with rerouting of passengers". In: *Transportation Science* 46.1 (2012), pp. 74–89.

[Hur12] E. VAN DER HURK, L. G. KROON, G. MAROTI, and P. H. M. VERVEST: "Dynamic Forecasting Model of Time Dependent Passenger Flows for Disruption Management". In: *Proceedings of CASPT.* Santiago de Chile, 2012.

[Man09] C. MANNINO and A. MASCIS: "Optimal real-time traffic control in metro stations". *Operations Research* 57.4 (2009), pp. 1026–1039.

[MAS02] A. MASCIS and D. PACCIARELLI: "Job shop scheduling with blocking and no-wait constraints". In: *European Journal of Operational Research* 143.3 (2002), pp. 498–517.

[Mes07] L. MESCHINI, G. GENTILE, and N. PAPOLA: "A frequency based transit model for dynamic traffic assignment to multimodal networks". In: *Proceedings of ISTTT17*. Ed by ALLSOP, BELL, HEYDECKER. Elsevier, 2007, pp. 407–436.

[Sels13] P. SELS, T. DEWILDE, D. CATTRYSSE, and P. VANSTEENWEGEN: "Expected Passenger Travel Time as Objective Function for Train Schedule Optimization". In: *Proceedings of the 5th IAROR conference*. Copenhagen, 2013.

[Sch07] A. SCHÖBEL: "Integer programming approaches for solving the delay management problem". In: *Robust and Online Large-Scale Optimization.* Vol. LNCS 4359. Springer, 2007, pp. 145–170.

[Suhl01] L. SUHL, C. BIEDERBICK, and N. KLIEWER: "Design of customer-oriented dispatching support for railways". In: *Computer-Aided Scheduling of Public Transport.* Ed. By: VOSS and DADUNA. Vol. *LNEMS* 505. Springer, 2001. pp. 365–386.

[Tam13] K. TAMURA, N. TOMII, and K. SATO: "An Optimal Rescheduling Algorithm from Passengers' Viewpoint Based on Mixed Integer Programming Formulation". In: *Proceedings of the 5th IAROR conference.* Copenhagen, 2013.

[Tor11] J. TORNQUIST KRASEMANN: "Design of an effective algorithm for fast response to the re-scheduling of railway traffic during disturbances". In: *Transportation Research Part C: Emerging Technologies* 20 (2011), pp. 62–78.

Corresponding author: Corman Francesco, Delft University of Technology, Mekelweg 2 26287CN Delft, The Netherland, e-mail:.f.corman@tu-delft.nl

Analyzing Railroad Congestion in a Dense Urban Network Through the Use of Road Traffic Network Fundamental Diagram Concept

Pierre-Antoine Cuniasse[1, 2], Christine Buisson[1], Joaquín Rodriguez[1], Emmanuel Teboul[2], David de Almeida[2]
[1]IFSTTAR
[2]SNCF

Abstract

Transilien, the branch of SNCF in charge of operating the main urban railroad network in the area of Paris, faces a regular increase of passengers flows. The planning of railway operations is made carefully: simulation runs permit to assess the timetable stability. However, many disturbance appear and cause trains delays. Due to the nature of the railroad network those delays are cumulative and an on-line update of the timetable is not always successful in maintaining the trains schedule. In this tensed context, operators are searching solutions to better use the infrastructure capacity and enhance the quality. A needed step towards this objective is a better understanding of the phenomena of disruptions. In particular because the expansion of congestion is not clearly understood until now. This paper explores the possibility to transpose a traffic flow theory tool, the network fundamental diagram, in the field of dense railroad traffic. Railroad traffic is different of road traffic by many ways: railways are a planned system, traffic volume does not satisfy the continuum hypothesis, stations force stops and the signalization system brings a discrete behavior. Despite those big differences we show how to build a network fundamental diagram for a railroad system and how to interpret some obtained shapes for those diagrams. These diagrams gives us some means to compare planned timetable and reality. We also identify the limits that need to be overcome to take benefits of the road traffic tools in railroad traffic analysis.

Keywords: Railroad, mass transit, congestion, Network Fundamental Diagram

1 Introduction

In dense urban regions, the railway system is frequently operating not as scheduled. Indeed, it is a common fact that theoretical schedules of trains are not respected and that in reality, the train circulation system is operating differently than what is planned. Deviations from schedules are observed in two cases: 1/ exceptional situations with the occurrence of an external event (due to meteorological conditions for example) 2/ everyday situations where the schedule is not strictly respected (for example a small delay in one station induces a bigger in the next one). Whatever the cause of the delay, railroad users are impacted and this is decreasing their confidence into railway mode, destroying therefore the efforts of modal shift to a more energy-efficient mode.

To face this problem, railway companies add margins to the theoretical schedule to increase the probability of schedule adherence; however, the added margins reduce the availability of track capacity for additional rail transport service. [UIC96; UIC04; Gov05]. Afterwards, the robustness of the obtained schedule is frequently evaluated with microscopic simulation tools [Nas04; Rad01]. Nevertheless, those increased efforts towards more realistic theoretical schedules are not effective. This is even truer if, like in dense urban regions, both political orientations toward sustainable transportation and the economic grow cumulate in an increasing demand.

Therefore, the lack of positive results for those approaches leads us to search for another intellectual scheme to reduce the differences between observed and planned schedules in urban dense railway operations. Analysis tools initially developed for other congested transportation modes can be considered and their transfer to the railway context has to be examined. A paper present the adaptation of Personal Car Equivalent concept to take into account the difference in mechanical abilities among trains in computing the capacity of a railroad [Lai13].

As a start, we consider a recently re-discovered tool for road traffic data analysis (the network fundamental diagram) and study its transfer to the case of the railway network of the Paris region, one of the most congested railway systems in Europe. The main aim of this paper is therefore to build the network fundamental diagram of railway lines (NFD-R) and to study the link between congestion and the shape of the NFD-R.

The paper is organized as follows. From a brief literature overview of road traffic network fundamental diagrams, we identify the main reported causes of dispersion of their points. A description of the railroad network and of the data collection technique is thus given. The results and their analysis is presented before a discussion. The paper ends with conclusions and recommendations for future research.

2 Network fundamental diagram : state of the art for road traffic networks

For more than half a century, road traffic is measured through local detectors (double electromagnetic loop detectors) that measure at a given point the time between the passing of two successive vehicles and the speed of each of them. Macroscopic variables characterizing the traffic flow seen as a whole can be extracted from those observations. Those variables are the spatial mean speed and the flow (number of vehicles passing the detector during a given period).

Expanding those measurements to compute mean accumulation and flow of vehicles for a given period over a whole network is not a new idea. After the initial publication of this idea in [God69], Herman and Prigogine propose in [Her79] a formulation of the relationship between mean speed and mean concentration. This idea was further developed in many papers of those authors and collaborators, where the idea was to examine the relationship between the mean flow on a road network and its total accumulation. Those papers are devoted to the free flow part of the diagram: the mean flow increases with the accumulation. In particular[Mah84] evokes the possibility of random events of lane blockages in a rectangular grid urban network simulated with NETSIM. The bigger the ratio between time with lane blockage and total simulation time is, the lower the slope of the relationship. Even with such lane blockages, it is worth noticing that the congested part of the relationship between global flow and global accumulation was not observed.

In [Dag07] Daganzo hypothesized that above a given threshold in total accumulation (critical value of accumulation), an increase of this accumulation leads to a decrease of the total flow. This was experimentally observed for the first time by Geroliminis and Daganzo one year after[Ger08; Ger07] for the congested center of Yokohama city. This relationship between total flow and total accumulation averaged during periods of typically a few minutes on a (sub-) network was initially named by Daganzo and Geroliminis Macroscopic Fundamental Diagram. Following the authors of [MSa12] in their introduction, we do prefer the term “Network fundamental diagram” which is to our opinion more precise, “macroscopic” referring to the variables (flow and concentration) by opposition with the microscopic ones like inter-vehicular time or individual speed.

Since 2008, network fundamental diagram exploration, has known many developments as a particular sub domain of road traffic flow studies. A recent and promising direction of research is the control of the most congested parts of a city network. It consists in splitting the city network into more than one reservoir and controlling the entry and exit flows of those reservoirs to decrease congestion (see for example [Had12; Key12]). Recently, one of these groups of authors proposed a method to partition the network into homogeneous zones to reduce the scatter of the diagram and increase the network controllability [Ji12].

Globally, as was first put into evidence in [Bui09] and further on mentioned in [Cas11], the NFD for a highway network is not well defined in the sense of [Ji12]. Indeed, the congestion level of the highway network is not homogeneous. In recently published papers, the impact of combining various traffic states among various parts of the considered network for computing the NFD has been explored in parallel by two teams: [MSa12] and also by [Dag11]. We will now present the transposition of the concept of Network Fundamental Diagram in the case of the dense and highly congested railroad network surrounding Paris.

3 Data and methods

3.1 Network

With a population of 11.5 millions inhabitants, Paris urban area is the most dense conglomeration in Europe. This density implies high levels of transportation demand and mass transit issues. For this paper we choose to focus on the SNCF (french national railroad) operated part of the RER (Local Express Network). Today, the growing demand makes the RER oversaturated. In this paper, we focus on the north part of B line and the central part of C line. The interest of those lines lies on their very high train traffic.

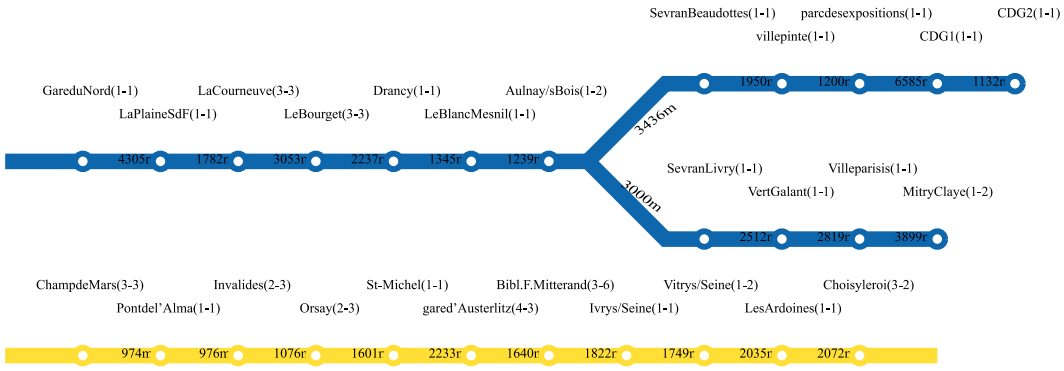


Figure 1: Analysed sections: the North-B line in blue and the central section of C line in yellow. This table provide the station name, the distance between consecutive stations, and in bracket the number of beacons in each station sector: $(\rightarrow, \leftarrow)$.

3.2 Data

Beacons are located over the railway network and detect the passage of trains and record their identifiers and events times. The beacon density over the network is variable and its order of magnitude is more than one beacon per station. Figure 1 is a description of the two analysed areas: the northbound of B line between Paris gare du nord, CDG Aeroport terminal 2 and Mitry Claye, and the central section of C line between Champ de Mars and Choisy le Roi. The results of this paper are based on a SNCF database which gathers all

the informations from the beacons and the corresponding theoretical times of passage over the beacons. Since each lines also got two directions, this gives four sets of data per day. Days with reduced demand (like week-end and holidays) or a special traffic event (planned public work...) are rejected. In the end, thirty-three days of data are available for B line and thirty-four for C line. The list of days used for each line is presented in table 2.

Figure 2: List of days selected for analysis. The red-bold days are for B and C lines and the green-italic ones for C line only.

4 Results and analysis

4.1 First observations of the network fundamental diagram for railroad

For one day and one set of data, the number of beacon events over time and the number of trains were counted and filtered with a running average (window time: 10 minutes, step time: 1 minute). We consider that at any given instant, the total number of trains present in a line is directly proportional to the average train concentration. We further assume that the number of trains passing over all the beacons of a line during a given period is proportional to an average flow on the same section. Note that this is not completely true in our case, since the distance between two consecutive beacons is not homogeneous over the line. Then, when plotting the number of events reported by all the beacons of a line against the number of trains of this line during the same period, we construct the NFD-R (Railroad) diagram.

Figure 3 presents the NFD-R for 2 lines: the B line - North direction (left column) and the C line - West direction (right column). To facilitate incident identification, markers were colored according to hours. The top diagrams of figure 3 represent the planned situation (T stands for Theoretical) on a typical day. The diagrams below present the results obtained in operation for various days. Depending on the day, a dispersion is observed or not. In any cases, most points are aligned along a line. We now examine the slope of the diagrams and, if present, the possible causes of spreading.

We can see from figure 3 that various shapes were observed for real diagrams. We visually selected days and lines where the realized NFD-R was almost linear (for example, diagrams (B-R2) and (C-R2) from figure 3 were rejected). This set of NFD-R is further referred as the “filtered NFD-R”. The numbers of filtered NFD-R are given in table 1. For those days and those lines, we made a proportional regression. Examples are presented in figure 4.

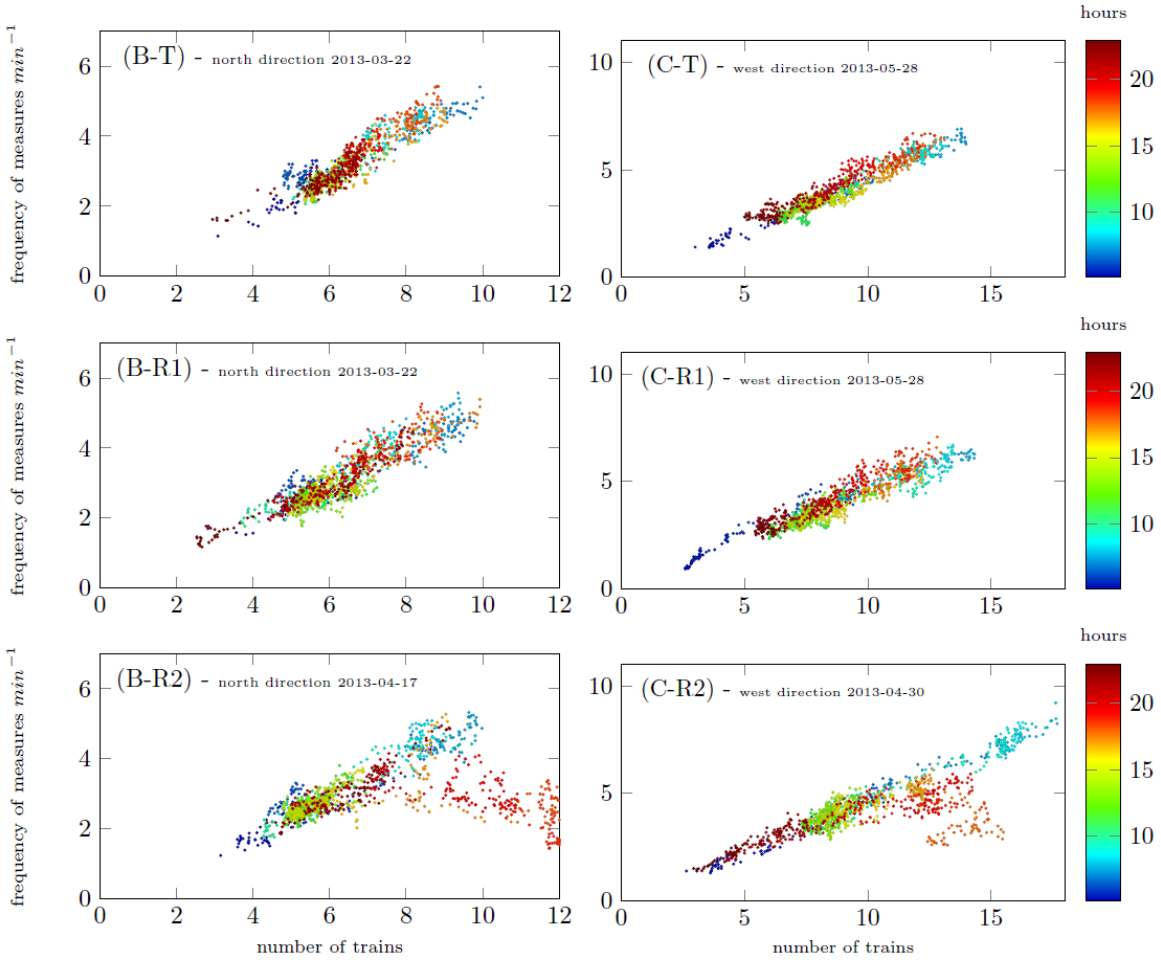


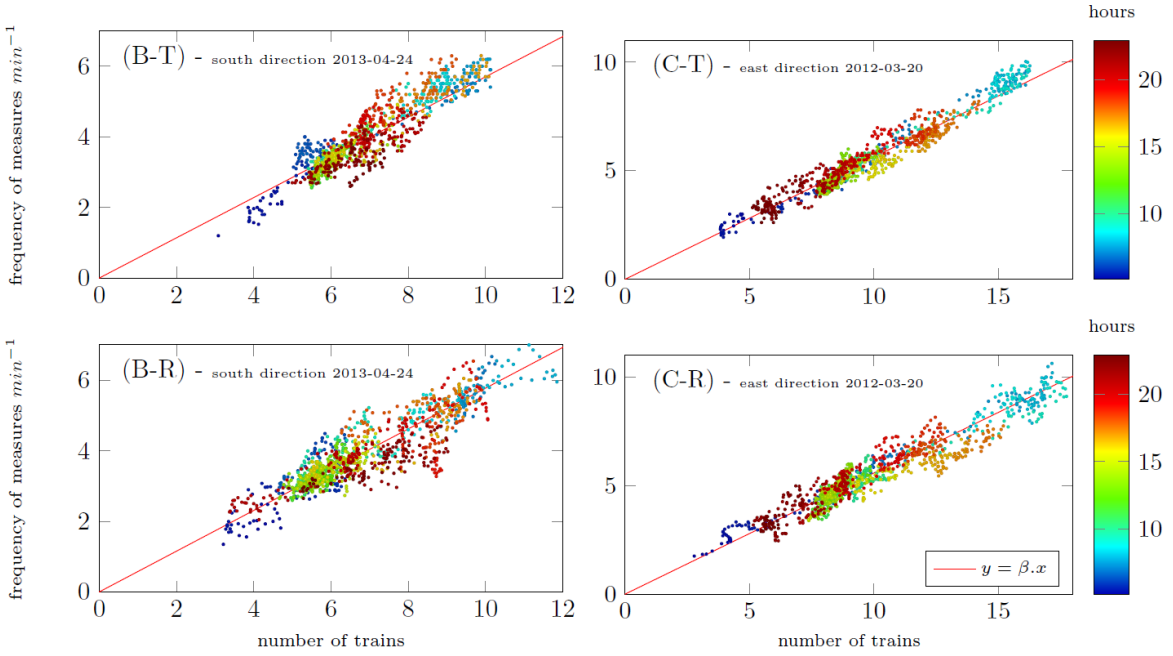
Figure 3: examples of NFD-R. (B-T) & (C-T) represent theoretical diagrams, (B-R1), (B-R2), (C-R1) and (C-R2) represent real diagrams.

Table 1: Average value for the residuals in different cases.

Line direction	B south T	B south R	B north T	B north R	C east T	C east R	C west T	C west R
Days	16	16	14	14	24	24	24	24
y_1	0.0116	0.0129	0.0101	0.0102	0.0117	0.0162	0.0090	0.0136

4.2 Comparison between theoretical and real NFD-R

It is well known that the slope of the left, uncongested part of a fundamental diagram is homogeneous to the speed in free flow conditions. To estimate a proxy for the railroad network, we use the β value which is homogeneous to a speed. β is then used to compare theoretical and realised NFD-R over all NFD-R. Figure 5 represents the frequency of the relative difference between the β values obtained for the actual and planned timetables ($\beta_{realised} - \beta_{theoretical}$) for the two studied lines. This figure shows that the trains of B line (resp. C line) travel often slower than planned. C line appears also more constant than B line. The lower size of the B line filtered sample compared to the C line might be an explanation for the higher spreading. The lower values of the slopes for the realized NFD-R compared



$y = \beta.x$	y		$y = \beta.x$	y	
(B-T)	$\beta = .570$	$res = .0113$	(C-T)	$\beta = .563$	$res = .0118$
(B-R)	$\beta = .577$	$res = .0131$	(C-R)	$\beta = .558$	$res = .0147$

Figure 4: examples of NFD-R, fitted with two mathematical relationship. (B-T) and (B-R): theoretical and real NFD-R for the B line in south direction. (C-T) and (C-R): theoretical and real NFD-R for the C line in east direction. $\beta = [min^{-1}.train^{-1}]$

to the theoretical NFD-R reflects a lower operating speed of the system, compared to what is planned. This has to be linked with the results obtained in simulations and presented in [Mah84], where the slope of the NFD of a urban road traffic network is lower when then duration of red phases of traffic lights is higher.

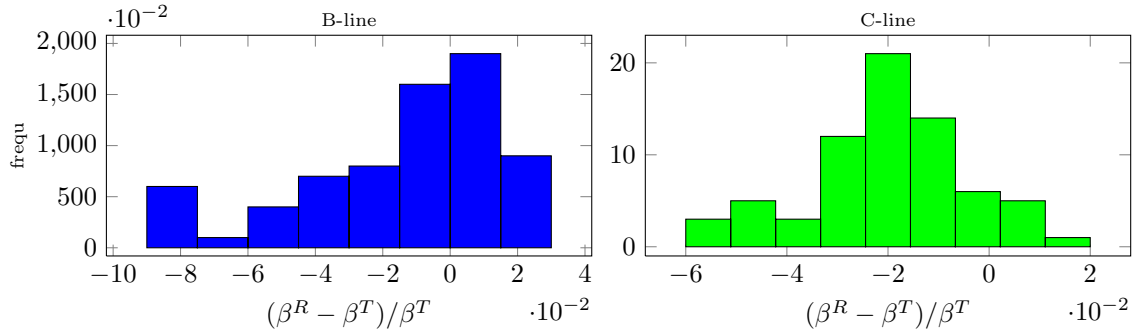


Figure 5: Comparison of β value fit for all NFD-R. For B line and for 60% of days $\beta_R < \beta_T$. For C line and for 89% of days $\beta_R < \beta_T$.

4.3 Impact of abnormal conditions of operations of the railroad network

The local fundamental diagram (FD) represents the traffic flow function of the concentration on a single point of a network. The figure 8.a presents the theoretical FD of operations of the trains for a network without stations. Traffic regulation over the French Network rely on blocking. The points U, C and FC of figure 8.a illustrate respectively three behaviours corresponding to three block aspects. Note that this FD is only theoretical for the moment. However, it can help us to construct theoretically a fictitious network fundamental diagram (see figure 8.b). In this NFD, the observed points are the result of a combination of various states of FD. If some trains do not operate in the free flow part of the FD, the NFD will not be linear. The figure 8.b illustrates this in the case of a combination of two different states of the network: points B and C of this figure are a combination of points b1 and b2 and c1 and c2 respectively. Logically, the point A, resulting of two free flow states, is located in the free flow part.

Nevertheless, in figure 4, for the case B-R2 the points located below the sloped line associated with congested operations. A first exploration has shown that in some cases, the occurrence of those abnormal points is linked in time with unplanned events. We have to better establish this point.

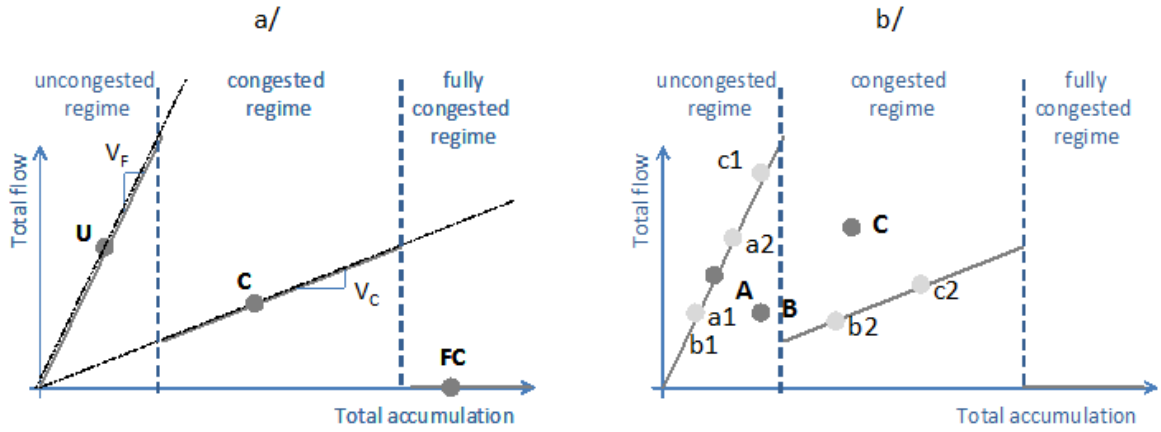


Figure 6: local (a) and network (b) theoretical fundamental diagrams of railroad operations in absence of stations. The bold lines of figure a presents the various equilibrium states associated with the train operations blocking rules.

5 Discussion and conclusion

We have shown in this paper that a network fundamental diagram (NFD-R) can be built for dense railroad Systems. This diagram is usually constructed for road traffic network both urban or freeways, and the literature is numerous where analysis of its shape is presented. Here we proposed a way to build it from identification by beacons of the passing of trains and we applied it to two lines where trains frequency is particularly high. With this diagram,

We highlight the existence of congestion phenomena. For most of the days, the slope of the NFD-R, which is an equivalent of the average speed, is lower than planned. This congestion phenomenon can appear in several ways: the first is discussed in this article and consist in a simple decrease of the speed while a NFD-R still exhibits a linear shape. On some of the NFD-R, on the contrary, we observe another shape, with dots below the sloped line. Those points will be analyzed in a further publication.

The research presented in this paper was only a first step of a larger project. The research directions to explore are numerous. Among them one can cite a better exploration of the data: a more precise study must be undertaken prior to eliminate the influence of variable beacons density and reinforce our results. An exploration of other lines will be needed to confirm the validity of our first results. Also the database of unplanned events has to be examined jointly with the points of the NFD-R located below the slope line. The impact of the stations and the accelerations they generate on the NFD-R must also be examined. Simulating at a large scale a railroad network with stations and realistic decelerations and acceleration will permit us to better understand the real NFD-R. We might also in a farer future explore through the NFD-R tool the feasibility of using cordon regulations to better cope with undesired congestion. The transposition of the recent findings of road traffic theory might help us in finding better ways of identifying and reducing congestion if the Paris suburban railroad network, provided that we keep in mind the significant differences among the two modes.

Acknowledgements

This research is conducted with the benefit of a PhD grant from Agence Nationale de la Recherche Technologique, France. Authors want to thank Ludovic Leclercq and Winnie Daamen for fruitful discussions during the preparation of this paper.

References

[Bui09] C. BUISSON and C. LADIER: "Exploring the Impact of the Homogeneity of Traffic Measurements on the Existence of Macroscopic Fundamental Diagrams". In: *Transportation Research Record* (2009).

[Cas11] M. CASSIDY, K. JANG, and C. DAGANZO: "Macroscopic Fundamental Diagrams for freeway networks : Theory and observation". In: *Transportation Research Record* 2260, 8-15 (2011).

[Dag07] C. DAGANZO: "Urban Gridlock: Macroscopic modeling and mitigation approaches". In: *Transportation Research B* 41(1), 49-62 (2007).

[Dag11] C. DAGANZO, V. GAYAH, and E. GONZALES: "Macroscopic relations of urban traffic variables: bifurcations, multivaluedness and instability". In: *Transportation Research B* 45(1), 278-288 (2011).

[Ger07] N. GEROLIMINIS and C. DAGANZO: "Macroscopic Modeling of Traffic in Cities". In: *Transportation Research Borad* No. 86 (2007).

[Ger08] N. GEROLIMINIS and C. DAGANZO: "Existance of Urban-Scale Macroscopic Fundamental Diagrams: Some Experimental Findings". In: *Transportation Research Part B* 42(9), 759-770 (2008).

[God69] J. GODFREY: "The Mechanism of a Road Network". In: *Traffic Engineering and Control* Vol. 11, No. 7, pp.323-327 (1969).

[Gov05] R. GOVERDE: "Punctuality of Railway Operations and Timetable Stability Analysis". PhD thesis. Delft University of Technology, 2005.

[Had12] J. HADDAD and N. GEROLIMINIS: "On the stability of traffic control in two-region urban cities". In: *Transportation Research B* 46(9), 1159-1176 (2012).

[Her79] R. HERMAN and I. PRIGOGINE: "A two-fluid approach to town traffic". In: *Science* 204, 148-15 (1979).

[Ji12] Y. JI and N. GEROLIMINIS: "On the spatial partitioning of urban transportation networks". In: *Transportation Research B* 46(10), 1639-1656 (2012).

[Key12] M. KEYVAN-EKBATANI, A. KOUVELAS, I. PAPAMICHAIL, and M. PAPAGEORGIOU: "Exploiting the fundamental diagram of urban networks for feedback-based gating". In: *Transportation Research B* 46(10), 1393-1403 (2012).

[Lai13] Y.-C. LAI, Y.-H. LIU, and Y.-J. LIN: "Development of Base Train Equivalents for Headway-Based Analytical Railway Capacity Analysis". In: *RailCopenhagen*. 2013.

[Mah84] H. MAHMASSANI, J. C. WILLIAMS, and R. HERMAN.: "Investigations of Network-Level Traffic Flow Relationships: Some Simulations Results". In: *Transportation Research Record 971* pp 121-130 (1984).

[MSa12] M. SABERI and H. MAHMASSANI: "Exploring the Properties of Network-wide Flow-Density Relations in a Freeway Network". In: *Transportation Research Record* No. 2315, pp. 153-163. (2012).

[Nas04] A. NASH and D. HURLIMANN: "Railroad simulation Using OpenTrack". In: *Proc. of the 9th COMPRail conference (Computers in Railways IX)*. 2004.

[Rad01] A. RADTKE and J.-P. BENDFELD: "Handling of railway operation problems with RailSys". In: *WCRR-Proceedings*. Köln, 2001.

[UIC04] UIC: "Capacity". In: *International Union of Railways* UIC code 406 (2004).

[UIC96] UIC: "Links between railway infrastructure capacity and the quality of operations". In: *International Union of Railways* UIC code 405 (1996).

Corresponding author: Pierre-Antoine Cuniasse, Transilien-SNCF, Paris, France, phone: +33 01 53 25 84 85, e-mail: pierre-antoine.cuniasse@sncf.fr

Calibrating and Validating Train Dynamics Characteristics against Realisation Data

Nikola Bešinović, Egidio Quaglietta, Rob M. P. Goverde
Delft University of Technology

Abstract

In the last decades advanced simulation models have been more and more used by railway timetable designers and dispatchers to support both the off-line planning and the real-time management of traffic. Fundamental requirements for these models are the accuracy and reliability of describing the train dynamics. Current models use default train behaviour that tend to significantly differ from the behaviour applied by drivers. To this aim it is necessary to estimate train running parameters against real data collected from the field. In this paper a simulation-based calibration approach is proposed to determine the parameters for the different phases of train motion (acceleration, deceleration, coasting and cruising) from track occupation data. A customized genetic algorithm is developed that minimizes the error between observed and simulated data. Model parameters are calibrated for different classes of trains against a significant number of real time-distance trajectories collected on the Dutch railway corridor Driebergen-Maarn. The model is validated by verifying the congruence of calibrated speed-distance profiles with those measured from the field with GPS devices. Results show that even in case of limited availability of track occupation data, our approach returns robustly calibrated parameters which accurately reproduce observed train behaviour. Reliable predictions are provided that can be used in practice to effectively support timetable planning and real-time traffic management.

Keywords: Train dynamics, running time models, simulation-based calibration, speed profile estimation, validation

1 Introduction

Traffic prediction models are increasingly spreading out and being used by practitioners in the railway field to support both off-line timetable design and real-time traffic management. The aim of both these activities is to generate a conflict-free schedule that allows the requested level of service availability also in presence of disturbed traffic conditions. The strategies that consent to achieve conflict-free train services are generally identified by the following three main steps: *i*) prediction of train trajectories, *ii*) detection of potential track conflicts (generated by trains that want to occupy the same block section at the same time),

iii) resolution of detected conflicts by solving a mathematical problem (e.g. optimization) and using measures such as retiming (i.e. delaying train departures), reordering (i.e. changing the order of trains at critical locations), rerouting (i.e. detouring trains on alternative routes), and/or speed adjustments. The effectiveness of computed schedules strongly depends on the reliability of the prediction model adopted to forecast the evolution of train trajectories. Indeed, only if train behaviour is described accurately it is possible to correctly detect possible conflicts and identify a suitable solution. To this purpose a proper calibration phase is needed to identify the parameters of the prediction model that allow to precisely describe real train dynamics. These model parameters must be fine-tuned against train data (i.e. speed and position) collected from the field in order to minimize the error between forecast and observed train paths.

In literature several authors [Alb06, Alb10, Med11] adopt different running time prediction models and calibration approaches. The main shortcomings of these works are that: *a*) [Alb06] and [Alb10] calibrate only the parameters of the kinematic motion equations which are trajectory-dependent and cannot be used anymore when considering a different train run even if the rolling stock is the same; *b*) [Med11] refer to the calibration of dynamic motion equations (e.g. Newton's motion law) but fine-tune only a single performance parameter for each phase of the motion, neglecting relevant inputs (e.g. coefficients of the tractive effort or the motion resistances) that strongly can influence the performance of the model; c) No validation phase is considered to verify the consistency of the calibrated model with observed train trajectories.

This paper presents a simulation-based model to estimate the parameters of the dynamic Newton's motion formula from real track occupation data collected from the field. Model parameters are calibrated for different classes of trains against a significant number of time-distance trajectories collected on the Dutch railway corridor Driebergen-Maarn. A genetic algorithm is developed to minimize the difference between observed and simulated trajectories. Validation of our approach is performed by verifying that speed-distance profiles derived from the calibrated model are congruent with those gathered from the field with GPS devices.

In the following the methodology and model used are described in Section 2. Section 3 reports the case-study adopted for the calibration experiment. Conclusions and final comments are given in Section 4.

2 Framework description

The calibration process is performed by means of a simulation-based framework that integrates a genetic algorithm (GA) with a microscopic running time model based on the dynamic motion equations of Newton [Han08]. Figure 1 illustrates the architecture and the components constituting the framework. Input of the running time model is represented by microscopic characteristics of the infrastructure (e.g. track length, gradients, speed limits, signals and station positions) and the rolling stock (e.g. train length), as well as the set of model parameters β to be calibrated. The vector β is composed of the coefficients of the resistance and tractive-effort equations as well as parameters which are specific for each phase of the motion: acceleration, cruising, coasting, and braking.

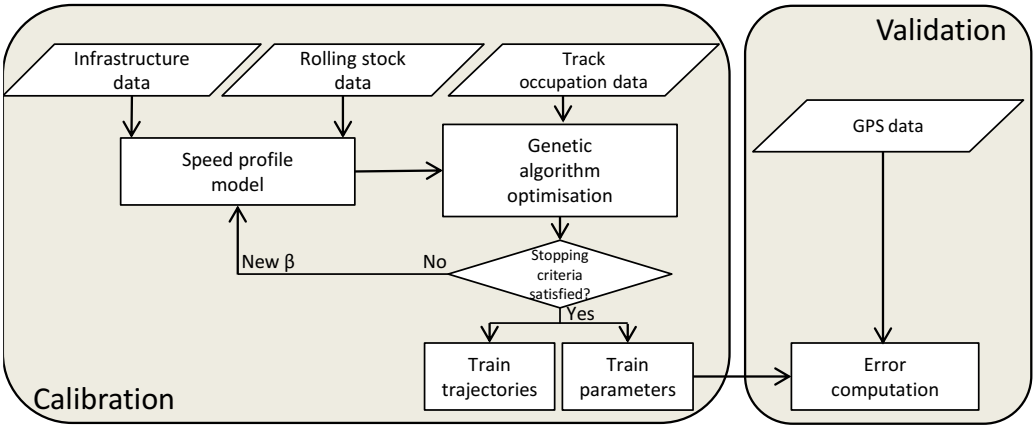


Figure 1: Functional scheme of the simulation-based optimization framework.

The value of each parameter of the vector β is randomly set by the GA at each iteration of the optimization and a corresponding train trajectory is computed by the running time model. This trajectory is then compared with the measured one from the field by means of track occupation data. Track occupation data are gathered by the Dutch train describer system TROTS [Kec12] which records for a given train (identified by an ID number) the times at which a certain track section is occupied/released. Only unhindered trains are considered (i.e. trains not running into restrictive signal aspects due to train path conflicts) since we are interested in understanding the unconstrained behaviour of train drivers when driving in free flow conditions. The objective function of the optimization problem is the absolute error between observed and simulated time-distance trajectories. If this error is too large, the algorithm will provide a new set of parameters and a new iteration is performed. Otherwise, the optimization process stops and the optimal set of parameters is returned as output together with the calibrated train trajectories (i.e. time-distance and speed-distance).

Our calibration process is then validated by verifying that calibrated speed-distance profiles are congruent with the real ones collected via GPS. The process is validated by quantifying the average error between simulated and measured GPS speed-distance profiles. In brief, we calibrate train parameters against time-distance trajectories observed from track occupancy data and successively validate the values with the speed-distance profiles collected by GPS. This framework is applied over a significant set of train runs for different train compositions (i.e. different multiple units sets) in order to estimate probability distributions for each of the parameters.

2.1 The microscopic running time model

The developed running time model is based on Newton's dynamic motion equation which is expressed as

$$f_t(v) - r(v) = f_s(v) = q \cdot m \cdot v \cdot dv/ds \quad (1)$$

Here, $f_s(v)$ is the surplus force used to accelerate the train obtained as the difference between the tractive effort, $f_t(v)$, and the resistance forces, $r(v)$, at speed v. Additional resistance produced by rotating train parts is expressed with rotating mass factor and denoted q. The tractive effort is generated by the traction unit and modelled as

$$f_t(v) = \begin{cases} c_0 + c_1 v, & v \le v_{overheat} \\ c_{i,k}/v, & v_k < v < v_{k+1}, i = 2..n \end{cases} \quad (2)$$

where the linear part of the function $(c_0 + c_1 v)$ is valid for values of speed lower than the overheat speed limit, $v_{overheat}$, while hyperbolic characteristics are valid for higher speeds representing a limitation due to adhesion and tractive power. It should be noted that engine characteristics may be defined with more than one successive hyperbolic curves [Han08].

The resistance forces are represented by

$$r(v) = r_0 + r_1 v + r_2 v^2 + f_C + f_G, \quad (3)$$

where the second-order polynomial of speed $(r_0 + r_1 v + r_2 v^2)$ depicts the sum of the resistances due to rolling and air viscosity, while f_C and f_G are the resistances due to curve alignment and track gradients. Coefficients r_0 and r_1 are relative to the mechanical resistance of the rolling stock, while r_2 refers to the aerodynamic resistance. The three coefficients also incorporate wind speed. The coefficients of both the tractive-effort and the resistance equations are considered in this research as mass specific, therefore equation (1) is divided by $q \cdot m$ and expressed in i.e., N/kg or kN/t.

Additional model parameters are considered to describe the cruising, coasting and braking phases of train motion. The parameter $\theta_{cruising}$ is used to define the cruising phase and represents the ratio between the speed limit on a track and the cruising speed actually operated. In other words, this parameter defines the compliance of the train driver to the track speed limit. The coasting phase is described by $\theta_{coasting}$ which is the ratio between the speed at the end and at the beginning of the coasting phase. The braking phase is depicted by b_{stop}, namely the braking rate adopted by the train driver.

Figure 2 illustrates the four train motion phases for a train running between two consecutive stops. As can be seen the running time model takes into account all characteristics of the network like static speed limits (dashed line), track gradients (lower black line), signal locations (purple circles) and dynamic speed limits provided by the Automatic Train Protection (ATP) system when approaching the arrival station (yellow rectangle).

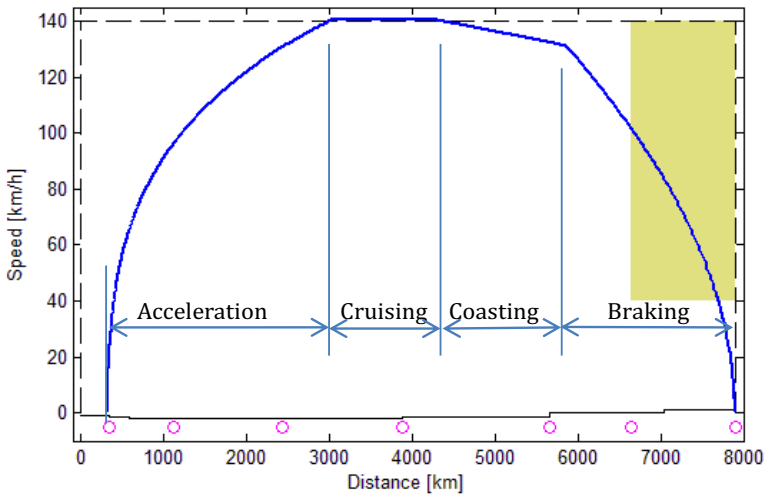


Figure 2: Phases of train motion and relative parameters.

2.2 The optimization model

The calibration process is formulated as an optimization problem that aims at minimizing the error between simulated and observed time-distance trajectories. Table 1 reports the parameters to calibrate, i.e. the decision variables of the optimization problem.

Table 1: Decision variables.

c_0	max starting tractive effort due to overheating limit[N/kg]
c_2	hyperbolic parameter of tractive effort function[Nm/s/kg]
r_0	constant resistance coefficient [N/kg]
r_2	quadratic resistance coefficient [Ns²/m²/kg]
b_{stop}	braking rate [m/s²]
$\theta_{coasting}$	coasting performance [%]
$\theta_{cruising}$	cruising performance [%]

The linear coefficients c_1 and r_1 are considered fixed and not calibrated since they are not sensitive parameters for the model as shown by previous analyses [Beš13a].

The mathematical formulation of the optimization problem is given as

$$Minimize \sum_{j \in N} |t_j^{observed} - t_j^{simulated}| \tag{4}$$

Such that

$$\frac{dv}{ds} = \frac{f_t(v) - r(v)}{v} \tag{5}$$

$$\frac{dt}{ds} = \frac{1}{v} \tag{6}$$

$$c_i \in \left[c_i^{lb}, c_i^{ub}\right], \text{ for } i = 0, 2 \tag{7}$$

$$r_i \in \left[r_i^{lb}, r_i^{ub}\right], \text{ for } i = 0, 2 \tag{8}$$

$$b_{stop} \in \left[b_{stop}^{lb}, b_{stop}^{ub}\right] \tag{9}$$

$$\theta_{cruising}(s) \in \left[\theta_{cruising}^{lb}(s), \theta_{cruising}^{ub}(s)\right] \tag{10}$$

$$\theta_{coasting}(s) \in \left[\theta_{coasting}^{lb}(s), \theta_{coasting}^{ub}(s)\right] \tag{11}$$

$$v(0) = v_0 = 0, \quad v(N) = v_{end} = 0. \tag{12}$$

The term $|t_j^{observed} - t_j^{simulated}|$ of equation (4) is the absolute error between the simulated ($t_j^{simulated}$) and observed ($t_j^{observed}$) passage time over the jth track section joint. The objective function is therefore the total absolute error computed over all N occupation data points collected from the field. At each iteration of the optimization algorithm the objective function is assessed by numerically integrating the speed and the running time as shown in (5) and (6). The integration process follows the adaptive Dormand-Prince method [But03] which is one from the Runge-Kutta family. Lower (lb) and upper (ub) bounds of each parameter are given in (7) - (11). Equation (12) gives the initial and final speed conditions : a train starts from a standstill and stops at the end of route.

A solution to the optimization problem is represented by the vector

$$\beta = (c_0, c_2, r_0, r_2, b_{stop}, \theta_{cruising}, \theta_{coasting}) \quad (13)$$

which contains the optimal value for each decision variable.

2.3 The customized genetic algorithm

A customized GA is developed in this research to solve the optimization problem. The algorithm works with a population of individuals (i.e. sets of train parameters β), each representing a possible solution. Each individual produces a different value of the objective function. The population evolves through generations towards better solutions (i.e. lower values of the objective function) by means of randomized processes of selection, crossover, and mutation (see [Mit96] for more information). The developed GA is customized to improve the performance of the algorithm according to the specific problem in such a way that computed individuals are stored in a list and not re-computed if proposed again within following generations. The algorithm runs in parallel to improve computing times of the optimization.

3 Case study: the Driebergen-Maarn corridor

The model parameters are calibrated against trajectories of trains running along the Dutch railway corridor Driebergen-Maarn, an eight kilometres long double-track line. The Dutch speed signalling system NS'54 with ATB automatic train protection [Alb10] is implemented over the corridor. Local, Intercity (IC) and international (ICE) trains operate on this line. The approach is applied only to the local trains since GPS data were available only for these ones.

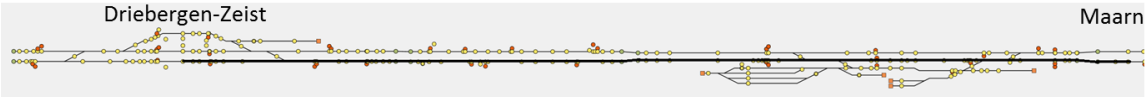


Figure 3: Schematic layout of the corridor Driebergen-Maarn.

Four different classes of local train compositions are analysed, but for brevity only the results relative to one of them is reported in this paper. Similar results were obtained for the others. The composition analysed is the Sprinter Light Train (SLT) consisting of six electrical units. The SLT6 has a length of 101 m and a maximum speed of 160 km/h, with the maximum track speed being 140 km/h. All observed trajectories used for the calibration have the same route, i.e., the same platform tracks, in-/outbound interlocked routes and block sections. Trains running in the Driebergen-Maarn direction are analysed.

For each train a total of N=10 track occupation points from TROTS are available, while GPS speed-distance measurements were collected each 5 s [She12]. Both TROTS and GPS data refer to the period September-November 2012. Although GPS data were initially processed, it may be considered that the noise due to the collection process and the accuracy of GPS device is still present to some extent and therefore might produce an additional error.

The calibration has been carried out on an AMD Athlon 3300 GHz processor with six cores and 4GB of RAM. The computation of a single train trajectory takes less than 10^{-2} seconds, while the computing time needed to complete a single calibration experiment is always lower than one minute.

3.1 Model calibration and validation

Results of the calibration are reported in Figure 4 for a single train. As can be seen, the calibrated time-distance trajectory (solid blue line) completely fits with the measured trajectory in the ten track occupation data points (red circles). The values of the train parameters calibrated for this train are also reported. As already done in [Beš13a, Beš13b], we have tested the stability of the optimization algorithm by repeating 30 times the calibration experiment for a given train and verifying the convergence of the calibration model.

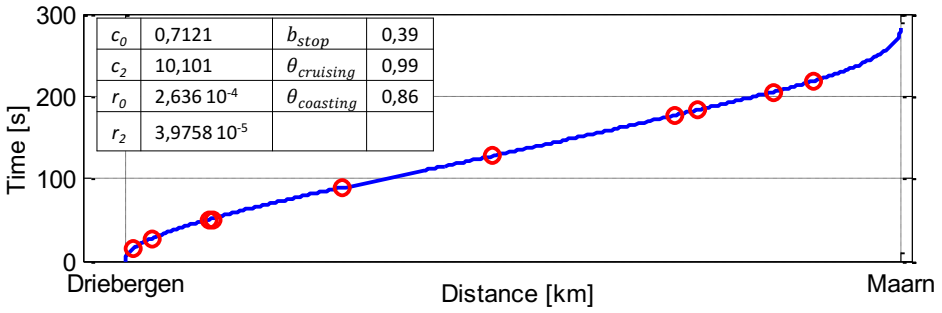


Figure 4: Estimated vs. nominal time-distance diagram for a single train run.

The train parameters of the SLT6 have been calibrated for 30 different observed train runs. Therefore, we are able to describe the statistical variation for each of the parameters due to the behaviour of train drivers when driving in free-flow conditions . Table 2 shows for each train parameter: the corresponding distribution function, the values of the distribution parameters and the KS-statistic of the Kolmogorov-Smirnov goodness-of-fit (the smaller it is the higher is the fit).

Table 3: Distributions of train parameters.

Parameter	Distribution Name	Distribution Parameters	KS-statistic
c_0	Weibull (3 parameters)	α=1,2792·10^{+8}; β=7,5841·10^{+6}; γ=-7,5841·10^{+6}	0,099
c_2	Generalized extreme value	k=-0,84333; σ=1,3676; μ=8,7108	0,133
r_0	Uniform	a=0,00625; b=0,00785	0,195
r_2	Generalized Pareto	k=-0,55525; σ=4,3414·10^{-5}; μ=1,5422·10^{-5}	0,057
b_{stop}	Jonhson	γ= -0.1337; δ= 0.6; λ=0.3146; ξ=0.244	0,073
$\theta_{cruising}$	Generalized extreme value	k=-0,77022; σ=0,02837; μ=0,98762	0,118
$\theta_{coasting}$	Johnson	γ = -0.7267; δ = 0.735; λ =0.289; ξ=0.721	0,065

For each train the mean absolute error (MAE_{time}) is computed:

$$MAE_{time} = \sum_{j=1}^{N} \frac{|t_j^{observed} - t_j^{simulated}|}{N} \qquad (14)$$

over the N=10 track occupation points. Table 3 presents the mean, the standard deviation and the maximum of the MAE_{time} over the 30 observed trajectories. The small values obtained by these metrics confirm that the calibrated model is able to reproduce reliable train trajectories in terms of running times.

Table 3: Errors between estimated and measured train trajectories (average over 30 trains).

		Error		
		Mean	Standard deviation	Max
Calibration	Time error [s]	0.42	0.25	1.24
Validation	Speed error [km/h]	2.48	0.81	7.24

However, if a model is reliable in assessing running times it does not necessarily imply that it is also reliable in estimating the speed profiles of trains. To this purpose the proposed calibration approach is validated by verifying that the parameters calibrated against track occupation data are also able to reproduce the real train speed profiles measured via GPS. The GPS speed-distance data collected for each one of the 30 train runs are illustrated in Figure 5.

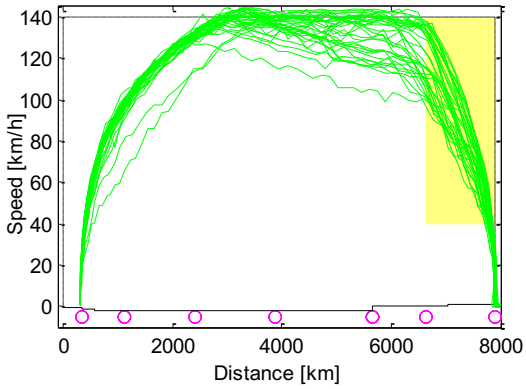


Figure 5: GPS train trajectories.

To quantify the deviation between simulated (given by the calibrated model) and real (GPS) speed profiles we calculate for each GPS point (measured each 5 s) the mean absolute error (MAE_{speed}):

$$MAE_{speed} = \sum_{j=1}^{N} \frac{|v_j^{observed} - v_j^{simulated}|}{N} \tag{15}$$

where $v_j^{observed}$ and $v_j^{simulated}$ are respectively the observed and simulated speed corresponding the jth GPS point.

The MAE are computed for each train, and the mean, standard deviation and maximum of the MAE over all 30 trains are reported in Table 3. The error of only 2.48 km/h implies a high reliability of the calibrated model in describing also the train speed profiles. This result is highlighted by Figure 6 that shows how close the simulated speed profile (solid line) is to the real one measured by the GPS (dotted line). The speed-distance profile is reconstructed well both for trains with (Figure 6*a*) and without (Figure 6*b*) coasting phase. For this latter case it is worth noticing the accuracy of the calibrated model to determine the slope, the beginning, and the end of the coasting phase.

Moreover, it is important to underline the ability of our approach in reliably describing both time-distance and speed-distance train profiles for calibrating parameters even when a small amount of track occupation data is available. This characteristic makes this approach suitable to be used also in corridors with limited availability of track occupation data.

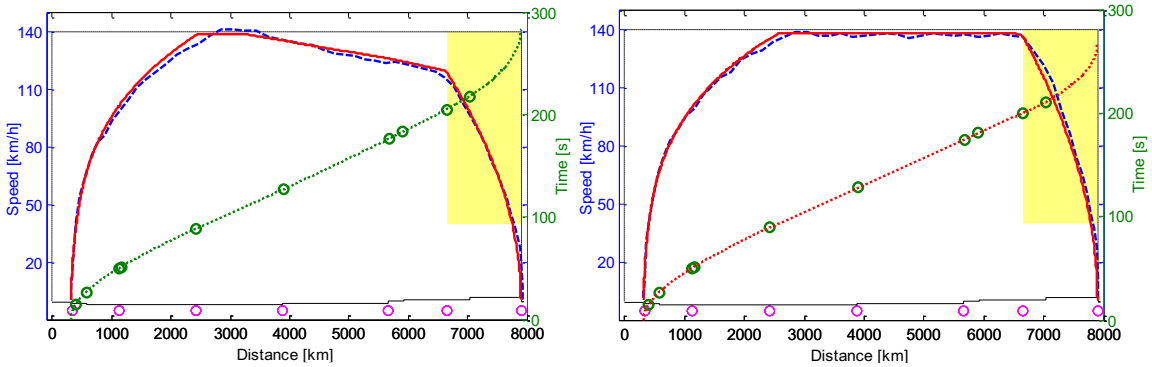


Figure 6: Estimated (solid) vs. GPS (dashed) dynamic speed profiles and time-distance trajectoryies (dotted) for a train: a) with coasting b) without coasting.

4 Conclusions and future work

This paper presented an approach to calibrate parameters of the dynamic motion equation of trains against track occupation data. A simulation-based optimization framework has been adopted to calibrate the coefficients of the tractive effort and the motion resistance equations, as well as specific parameters for each phase of the motion, i.e., braking, cruising, and coasting. Model parameters are calibrated for different classes of compositions for local trains running on the Dutch railway corridor Driebergen-Maarn. Calibrating the parameters against a significant set of observed train runs has consented to identify the statistical variation for each parameter due to the unconstrained behaviour of the train driver when driving in free-flow regime. Moreover, the small deviation between simulated and observed time-distance trajectories confirms the accuracy of the calibrated model in estimating train running times. A validation phase was realized to verify the congruency of the simulated speed-distance profile versus the real one measured from the field via GPS. The small average error between simulated and measured speeds underline that the calibrated model returns reliable train speed profiles as well. This conclusion is valid for all observed trains independently from the fact that they present coasting phases or not. In particular, the model estimates remarkably well the slope, the beginning, and the ending points of the coasting phases.

The proposed approach reproduces train behaviour also if calibrated against the measurement points of track occupation data. It has been shown that the model is accurate even when a small number of measurement points is provided. This makes it suitable to real-world applications also with corridors having a limited availability of these kind of data.

Future research will be addressed to understand how train parameters vary between off-peak and peak train runs, as well as between delayed versus non-delayed trains. Moreover, the method will also be extended to estimate the speed profiles of hindered trains.

Acknowledgement

This research is funded by the EU FP7 project “Optimal Networks for Train Integration Management across Europe” (ON-TIME). We also thank the Dutch infrastructure manager

ProRail and Pavle Kecman for providing the track occupation data, and the Netherlands Railways (NS) and Gerben Scheepmaker for the GPS data.

References

[Alb06] T. ALBRECHT, R. M. P. GOVERDE, V. A. WEEDA, and J. VAN LUIPEN: "Reconstruction Of Train Trajectories From Track Occupation Data To Determine The Effects Of A Driver Information System". In: J. ALLAN, C. A. BREBBIA, A. F. RUMSEY, G. SCIUTTO, S. SONE, (EDS.): *Computers In Railways X.* Southampton: Wit Press, 2006, pp. 207-216.

[Alb10] T. ALBRECHT, C. GASSEL, A. BINDER, and J. VAN LUIPEN: "Dealing With Operational Constraints In Energy Efficient Driving". In: *Proceedings Of The 4th Iet International Conference On Railway Traction Systems (Rts 2010).* Birmingham, 2010.

[Beš13a] N. BEŠINOVIĆ, E. QUAGLIETTA, and R. M. P. GOVERDE: "A Simulation-Based Optimization Approach For The Calibration Of Dynamic Train Speed Profiles". In: *5th International Seminar On Railway Operations Modelling And Analysis (Railcopenhagen 2013)*. Copenhagen, 2013.

[Beš13b] N. BEŠINOVIĆ, E. QUAGLIETTA, and R. M. P. GOVERDE: "Estimating Dynamic Train Running Time Models Against Track Occupation Data Using Simulation-Based Optimization". In: *16th IEEE Conference On Intelligent Transportation Systems (ITSC 2013)*. The Hague, 2013.

[But03] J. C. BUTCHER: *Numerical Methods For Ordinary Differential Equations*. London: Wiley, 2003.

[Han08] I. A. HANSEN and J. PACHL: *Railway Timetable And Traffic*. Hamburg: Eurailpress, 2008.

[Kec12] P. KECMAN and R. M. P. GOVERDE: "Process Mining Of Train Describer Event Data And Automatic Conflict Identification". In: C. A. BREBBIA, N. TOMII, P. TZIEROPOULOS, J. M. MERA (EDS.): *Computers In Railways XIII.* Southampton: Wit Press, 2012, pp. 227-238.

[Med11] G. MEDEOSSI, G. LONGO, and S. DE FABRIS: "A Method For Using Stochastic Blocking Times To Improve Timetable Planning". In: *Journal Of Rail Transport Planning & Management* 1.1 (2011), pp. 1-13.

[Mit96] M. MITCHELL: *An Introduction To Genetic Algorithms*. Cambridge, Ma: Mit Press, 1996.

[Sch13] G. SCHEEPMAKER: "Rijtijdspeling In Treindienstregelingen: Energiezuinig Rijden Versus Robuustheid". Master Thesis, Delft: Delft University Of Technology, 2012.

Corresponding author: Nikola Bešinović, Department of Transport and Planning, Delft University of Technology, P.O. Box 5048, 2600GA Delft, The Netherlands, phone: +31 15 2784914 , email: n.besinovic@tudelft.nl

Calibration of a Data-driven Railway Traffic Prediction Model

Pavle Kecman, Rob M. P. Goverde
Delft University of Technology, Department of Transport & Planning

Abstract

Monitoring and traffic state prediction are important tasks of railway traffic controllers. Recently developed prediction model gives accurate predictions of departure and arrival times, and route and connection conflicts for all trains in the controlled area. The process times of the graph model are learned from the historical data extracted from train describer log files. In this paper we test and validate the assumptions that running and dwell times depend on actual delays, analyse the usage of running time supplements and determine the time loss resulting from route conflicts. Findings from the data analysis are built in the data-driven prediction model. The model is validated in a simulated real-time environment on a real-life case study of a busy corridor between Leiden and Dordrecht in the Netherlands.

Keywords: Railway traffic, Track occupation data, Data analysis, Prediction

1 Introduction

Railway traffic controllers have a complex task to keep track of current train positions on their part of the network, predict the future evolution of traffic, and provide a set of control actions that will reduce the deviations of train paths from the planned schedule [Lue09]. In current practice, only the last measured train delays are known in the traffic control centers and dispatchers must predict the arrival times of trains using experience only, without adequate computer support. This often results in simple extrapolation of the current delays for the expected arrival delays. This method neglects the fact that some trains may (partially) recover from a delay using running time supplements, while others may get (more) delayed due to route conflicts.

Variability of process times and its influence on reliability of railway schedules has been studied intensively in recent years. Medeossi *et al.* [Med11] used track occupation data along with train event data recorded on-board to calibrate the train motion parameters in the process of computing stochastic blocking times for individual trains. In another approach, Büker & Seybold [Bue12] modelled delays as random variables, described with suitable

distribution functions, and applied analytical methods to compute delay propagation in a mesoscopic graph-based model. The large-scale character of the models does not allow precise modelling of train interactions and the resulting variability in running times.

Another stream of research was directed to short-term predictions of train traffic in real-time. Microscopic simulation tools such as *OpenTrack* [Nas04] or *RailSys* [Sie06] are able to give accurate predictions of running times, and possible route conflicts and resulting delay propagation. Due to a high level of detail in modelling infrastructure and train dynamics, such models are not suitable for real-time applications on large and heavily utilised networks.

Hansen *et al.* [Han10] presented a macroscopic model for prediction of train running times using historical track occupation data. An on-line prediction tool based on a directed acyclic graph with arc weights that are computed using train motion equations has been implemented in the Swiss traffic control system RCS-DISPO [Dol09]. Prediction errors smaller than 1 minute were obtained for events within 20 minutes prediction horizon.

In an earlier work, Kecman & Goverde [Kec13b] presented an approach based on computing arc weights of a microscopic graph model using historical data. A depth-first search based algorithm for computing the predicted event times over a graph with dynamic arc weights gives predictions for all reachable events within the horizon. This approach was extended [Kec13a] by precise modeling of route conflicts and incorporating time losses, due to braking and re-accelerating of hindered trains, in the predictions. Moreover, an adaptive component that exploits the feedback information about the actually realized blocking times of running trains has been implemented.

This paper focuses on analysing process times using historical track occupation data. By relying on actually realized processed times, learned from the data, rather than on theoretical values of process times, the prediction of event times for an individual train captures the phenomena of train behaviour. Running and dwell times as well as minimum headway times and route conflicts were analysed from historical track occupation data. The findings are incorporated in a dynamic computation of graph weights for a traffic state prediction model.

After describing the prediction model (Section 2), methodology and results of data analyis aare given in Section 3. Results form a real-life case study are presented in Section 4 and conclusions and directions for future research in Section 5.

2 The Online Traffic Prediction Model

The online prediction model is based on a directed acyclic graph with dynamic arc weights [Kec13b]. The graph topology is built and updated based on the actual train orders, route and connection plan. We distinguish between signal events (passing of a signal by a running train) and station events (arrival/departure at at/from a platform track). Events of a train are connected by running and dwelling arcs. Interactions between trains are modeled by headway and connection arcs. The graph is constructed in a way that fully reflects the microscopic operational constraints of railway traffic, described by blocking time theory on open track segments (between two stations) and the route setting and release principle in

station areas [Han08].

We assume that the actual route and connection plans are continuously provided by the traffic control system for the period of the prediction horizon. Each change of the actual plans or new information from the real-time operations results in an update of the graph topology, i.e., adding new trains, modifying train routes, updating connections and removing passed events. The predictive model can also be used to evaluate the effect of potential dispatching actions before implementing them in the form of a working timetable.

Arc weights represent the predicted process times for running and dwelling arcs and minimum process times for headway and connection arcs. The weight of a running or dwell arc is time-dependent and assigned in a dynamic way, depending on the (estimated) starting time of the modeled process. Arc weights depend on the actual delays (difference between realized and scheduled event times) and predicted delays (difference between predicted and scheduled event times). That way the dependence of running and dwell times on current (predicted) delays is incorporated in the model. Moreover, the graph weights are adjusted dynamically to incorporate the effects of predicted route conflicts and to minimize the prediction error for running trains based on the already realized running times [Kec13a].

3 Analysis of track occupation data

Track occupation data, obtained by processing the train describer log files of the Dutch train describer system *TROTS* (*Train Observation and Tracking System*) [Kec12], are used to calibrate the prediction tool with actually realized rather than theoretical process times. Three months of data from a busy corridor between Leiden and Drodrecht in the Netherlands were split into a training set containing 80 days and test set of 10 days of track occupation data. Route conflicts are identified and only conflict-free process times are used in the analysis of process times and model calibration.

The weights of running and dwell arcs are assigned dynamically [Kec13b]. The main idea behind this approach is that the running and dwell time of a train depends on the current delay. Delayed trains may run with full performance in order to use the running time supplements to reduce the delay. On the other hand, trains running on time or ahead of their schedule run with in lower performance regime, thus avoiding early arrivals and achieving energy efficient driving.

Similarly, the dwell times of trains in stations may depend on arrival delay. Since trains cannot depart from a station before the scheduled departure time, early trains have longer dwell times than scheduled in order to avoid early departures. On the other hand, trains with a positive arrival delay spend minimum dwell time in order to minimize the departure delay.

These general assumptions were tested on the data set of track occupation data. Track occupation data of each train line were analysed separately, thus ensuring that the stopping pattern and routes of all observed trains are the same. Correlations between running and dwell times with actual delays are tested using least trimmed squares (LTS) robust linear regression resisting 25% of outliers [Rou06].

The assumption about the different behaviour of delayed (delay larger than 60 seconds) and punctual trains was tested by separating the set of observed running and dwell times into corresponding sets of delayed and punctual trains and applying the Wilcoxon rank sum test at 5% significance level. The null hypothesis is that samples have continuous distributions with equal medians.

3.1 Running Times

Observed running times over a section between two stations show no correlation ($R^2 = 0.0012$) with departure delay. Figure 1 (left) shows observed running times of train line 2200 between The Hague HS and Delft. The red and black dashed line represent the robust fit and the 10^{th} percentile of running times, respectively. Scheduled running time is 420 seconds.

The null hypothesis was accepted for this data set ($p = 0.4028$) indicating that no significant difference in running time distributions for delayed and punctual trains was determined. The right side of Figure 1 shows the box-plots of running times for different delay values. Small variation of running times depending on departure delay can be observed.

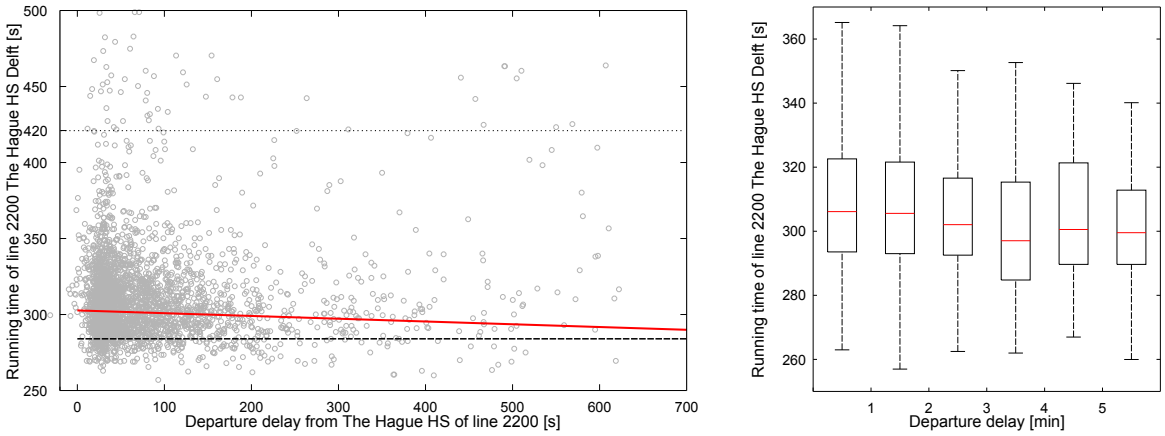


Figure 1: Dependence of running time on delay (left), box-plots of running times for punctual and delayed trains (right).

We expected the difference in performance regimes to be reflected to the greatest extent at the last part of the section before the scheduled stop where punctual or early trains have longer running times due to coasting or cruising with lower speed. The realised running times were presented relative to the departure delay. Weak correlation between running times and departure delays was found on the level of block sections ($R^2 = 0.0376$). This is illustrated in Figure 2 (left) which shows the dependence of running time over the last block before the scheduled stop in stations Delft of train line 2200.

The Wilcoxon rank sum test rejected the null hypothesis with $p \approx 0$. Box-plots in Figure 2 (right) show small differences in distributions of six data samples specified based on the value of departure delay.

Figures 1 and 2 show that delayed trains indeed tend to run faster to recover the departure delay. However, for punctual, early or slightly delayed train no such correlation can be

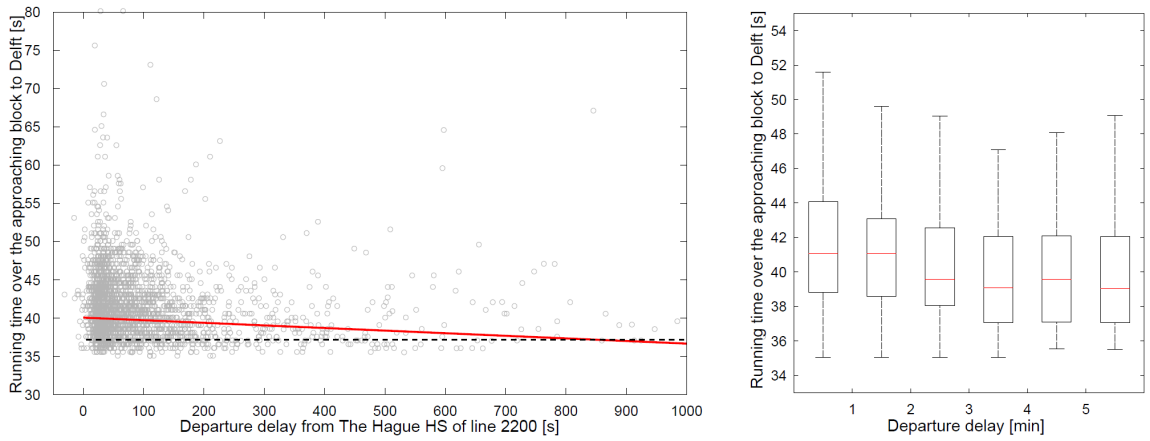


Figure 2: Dependence of running time on delay (left), box-plots of running times for punctual and delayed trains (right).

established. Earlier analyses of running times over sections between two scheduled stops, conducted in the Netherlands [Han10] and Switzerland [Lue09] show similar results.

Since no or weak correlation between running times and actual delays was discovered, it is important to determine how the running time supplements are actually used. In order to do so, delay accumulation over all scheduled stops for each line was analysed. Figure 3 shows how the delay of line 2200 trains changes over the route along the corridor Leiden - Dordrecht. The red line indicates the mean of delay change over space. No distinction can be made between early, punctual and delayed trains. It is visible that time reserves are spent on extended dwell times. Trains generally run full performance thus compensating for departure delay (delayed trains) or having more slack during dwell times (punctual trains).

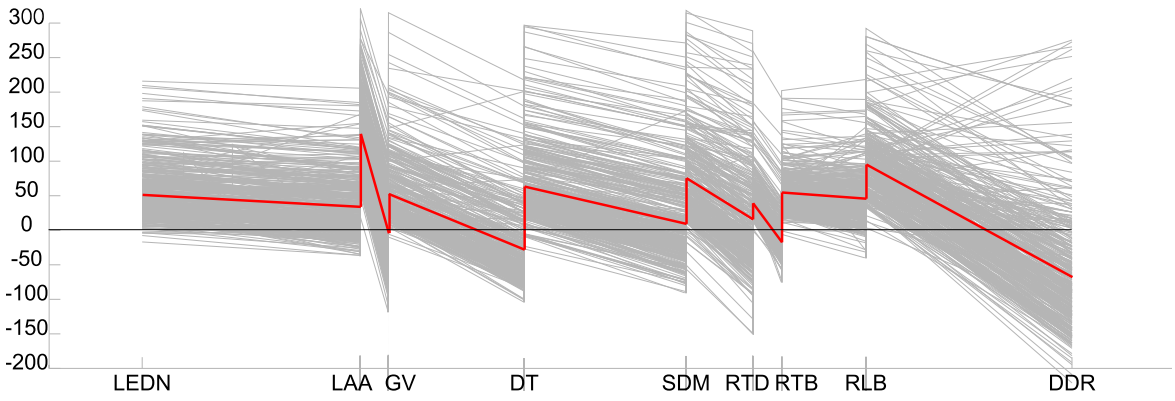


Figure 3: Delay over corridor Leiden - Dordrecht for train line 2200.

We emphasise here that the methodology for deriving the arrival and departure times relies solely on train describer data, thus an error of up to 10 seconds can occur depending on the topology of track circuits [Kec12].

3.2 Dwell Times

Availability of data from door sensors and on board equipment has inspired recent research in detailed modelling of train dwell times [Med11]. In this paper we rely solely on train describer data, thus detailed analysis of different phases of dwelling in scheduled stops was not possible [Kec12].

Dependence of dwell times on arrival delays was examined. Figure 4 (left) shows the dependence of dwell times on arrival delays for the train line 2200 in station Delft. Horizontal black dashed line represents the 10^{th} percentile of all dwell times, whereas the red line represents the robust linear fit for punctual trains. Scheduled dwell time is 60 seconds. Strong correlation ($R^2 = 0.8704$) was captured for early and punctual trains. Wilcoxon rank sum test rejected the null hypothesis ($p \approx 0$) and different distributions of dwell times for punctual and late trains are clear from the box-plots in Figure 4 (right). However, variation of dwell times for delayed trains needs to be explained by other factors and therefore, the data set is divided into a set of punctual and delayed trains.

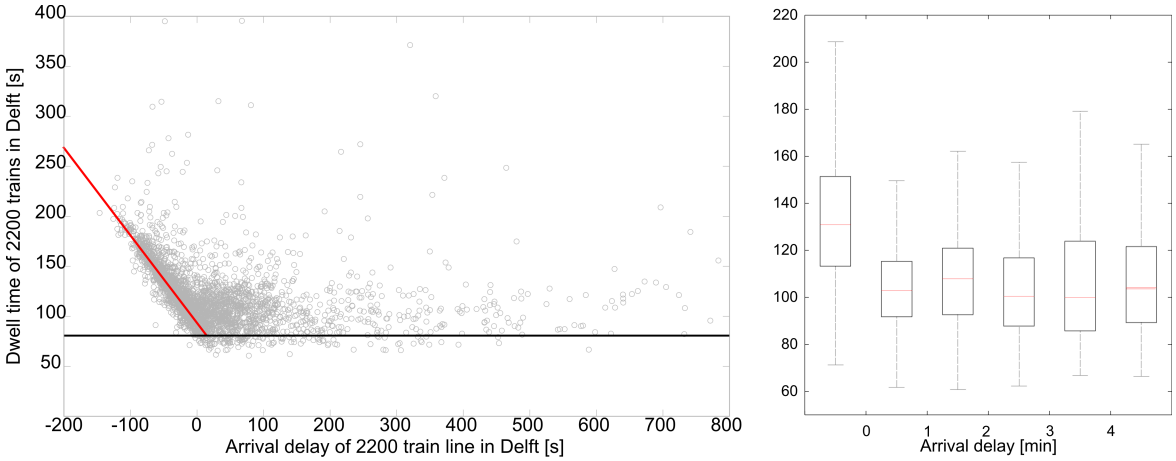


Figure 4: Dependence of dwell time on delay (left), box-plots of dwell times (right).

Variability of dwell times of delayed trains is explained by dependence on the time of the day. Dwell times of delayed trains normally equal the minimum dwell time required for passenger operations and route setting. We assumed that passenger volumes and consequently time needed for boarding and alighting increases during peak-hours. Figure 5 shows dwell times (weekends and holidays were not considered) relative to scheduled arrival times of the train line 2200 in Delft. The increase in dwell times during peak-hours is clearly visible. The red line indicates the median dwell time.

This clear distinction between causes of variability of dwell time for punctual and delayed trains requires a bimodal approach to prediction of dwell times. Therefore, for punctual and early trains, dwell time can be predicted based on the correlation with arrival delay. On the other hand, dwell time for a delayed train will be estimated from historical data based on dwell times of the same train number and adjacent train numbers of the same series (e.g. if train 2245 arrived with a delay, the dwell time will be predicted as the average dwell time of

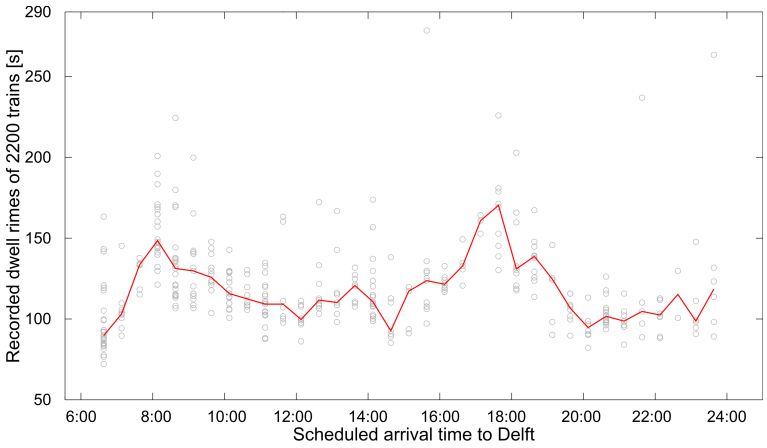


Figure 5: Dependence of dwell time on scheduled departure time.

trains 2243,2245 and 2247 obtained from the data set of delayed trains).

3.3 Headway and Connection Times

The weight of a headway arc represents the minimum time from the moment when the head of the first train leaves a block section to the moment when the next train can occupy the same block. Minimum headway time equals the sum of block clearing time by the first train, and setup and release time of the signalling system [Han08]. In this paper a constant value of 2 seconds is used for the setup and release time on open track and 12 seconds for route setting time in stations. Clearing time is estimated from the data as the 10^{th} percentile of the clearing times of a block by a specific train line.

In order to model the principle of sectional release using only signal passing events, the minimum headway time between two trains with diverging or intersecting routes is estimated from the data as the 10^{th} percentile of the time headways between train runs of the corresponding train lines from the historical track occupation data. By choosing a small percentile of the realised time headways, the impact of buffer times on minimum headway times estimates is excluded.

The weight of a connection arc is equal to the minimum transfer time for passenger connections or the time needed to perform activities that enable planned rolling-stock and crew circulations, for logistic connections.

3.4 Route Conflicts Analysis

The impact of a route conflict on the running time of the hindered train over the subsequent block depends on the conflict duration and the route and running time of the hindering train. The typical situation that occurs in practice when the two conflicting trains follow the same route is the ‘conflict wave’, where the hindered train keeps passing signals that show yellow aspect and is thus unable to re-accelerate to full speed [Gov11]. We therefore consider the time loss due to re-acceleration only after the hindered train passes a green aspect signal.

Since the running time estimates are computed based on the free running times, arc weights that model the running processes over affected blocks need to be adjusted to take into account braking (and possible waiting time in front of the signal), running at a lower speed and re-acceleration for every predicted route conflict.

In order to estimate the effects of route conflicts on train running times, all route conflicts in three months of traffic on the busy corridor Leiden–Dordrecht in the Netherlands were filtered out. The impact of the conflict duration on time loss after passing a yellow signal was analyzed. The duration of a conflict is computed using the process mining conflict identification tool [Kec12] and the resulting time loss is obtained as a difference between the realized running time over a block and the predicted conflict-free running time derived depending on the current train delay.

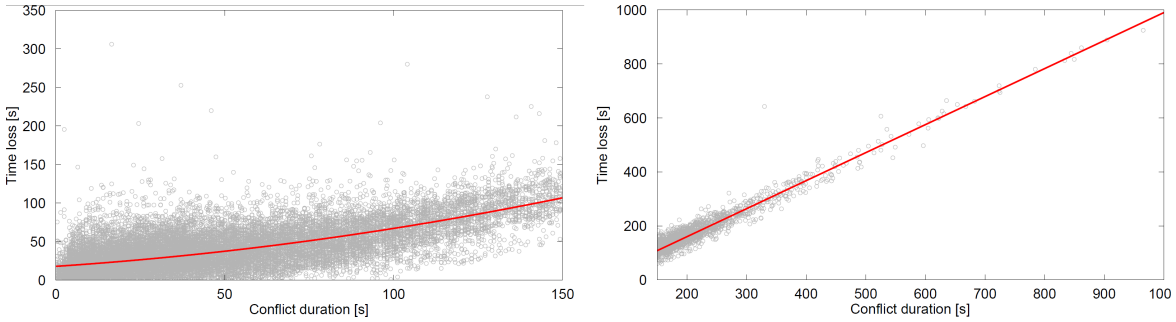


Figure 6: Dependence of time loss on conflict duration for conflicts shorter than 150 s (left) and longer than 150 s (right)

A regression analysis was performed based on 20130 data points split into points for conflicts shorter than 150 seconds (Figure 6 left) and points for conflicts longer than 150 seconds (Figure 6 right). A robust quadratic fit resisting 25% of the outliers showed the best performance in terms of coefficient of determination $R^2 = 0.79$ for conflicts shorter than 150 seconds. Even though the data points are scarce for conflict duration longer than 150 seconds, the slope of the linear regression line ($R^2 = 0.92$) can be interpreted easily as the waiting time in front of the signal and intercept as the time loss due to braking to standstill.

The time loss due to re-acceleration (after passing the green aspect signal) was also analyzed but no correlation with conflict duration was found. This can be explained by the fact that a train starts re-accelerating before it passes the green signal aspect, independent of the conflict duration at the previous signal.

The prediction algorithm [Kec13b] identifies predicted route conflicts and their duration. Running times of hindered trains are than adjusted to incorporate the time loss [Kec13a].

4 Case Study

The predictive model has been tested and validated on the busy corridor Leiden–The Hague–Rotterdam–Dordrecht in the Netherlands. The 60 km long corridor is (partially) traversed

daily by approximately 300 trains per direction.

For model validation (and example of application) we simulate the real-time environment by scanning the train describer log file from the test set that contains the chronologically sorted infrastructure and train messages from the train describer log files of two traffic control areas (Rotterdam and The Hague) for one day of traffic. Traffic control input is included in the form of a list of trains described by train number, timetable, route plan (block sections) and expected entrance time to the observed part of the network (or the first departure times if the train starts within the observed area).

4.1 Comprehensive evaluation

We tested the model performance by sweeping the test set train describer log file with rolling prediction horizons of different length. The prediction algorithm is initiated and the rolling horizon is moved after receiving a message that reports the realization of each of the 9776 signal and station events during one day of traffic on the corridor. Table 1 shows the average absolute prediction error, standard deviation, the average number of events that are predicted in each algorithm execution, and the average number of arcs for prediction horizons of 2 hours, 1 hour, 30 minutes, 20 minutes and 10 minutes.

Table 1: Model performance for different prediction horizons

	Prediction horizon [min]				
	120	60	30	20	10
Average error [s]	58.69	53.80	46.56	39.72	24.07
Standard deviation [s]	94.21	93.55	84.07	76.76	57.96
Average no. events	1040	532	269	180	90
Average no. arcs	2288	1117	590	389	202

Since the prediction algorithm is linear, computational complexity, which depends on the size of the input graph is not considered as a criterion for choosing the most appropriate prediction horizon. Even for the longest prediction horizon, the algorithm execution takes less than one second. Average prediction error, as well as the average number of nodes and arcs are monotonically decreasing as shorter prediction horizons are considered.

Figure 7 (left) shows box-plots of errors of event time predictions for each considered prediction horizon. Standard deviation of prediction error reduces with the decrease of horizon length. The right side of the figure presents distribution of prediction error of process times. It is clear that dwell times are the major sources of inaccuracy that are further propagated through the graph. Therefore, dwell times need to be modelled with higher precision since the variation of prediction error is significantly larger than for running times.

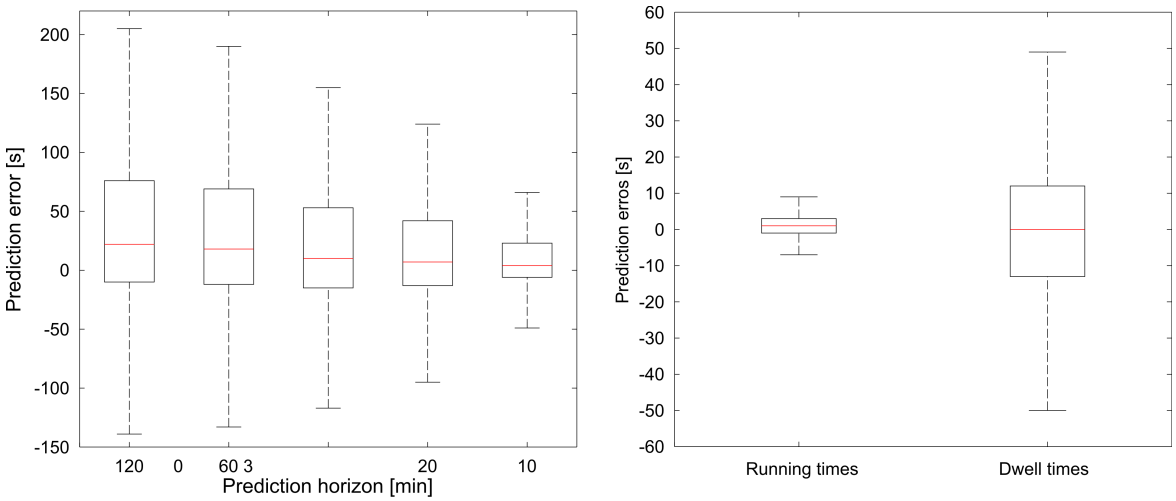


Figure 7: Box-plot of prediction errors for different horizon length (left) and prediction error for process times (right)

5 Summary and conclusions

This paper presented a data driven approach for calibrating the model for traffic prediction. The model has been applied in a real-life case study on a busy corridor in the Netherlands in a simulated real-time environment, and produced accurate estimates for train traffic and route conflicts within prediction horizon.

Weak dependence on actual delays has been established for running times. The analysis showed that the majority of trains run in full performance regime regardless of departure delays. Furthermore, clear bimodal behaviour of dwell times was captured. Dwell times of punctual trains show strong correlation with arrival delays, whereas delayed trains are more sensitive to impact of peak hours. Time loss due to route conflicts depend on the conflict duration. Bimodal behaviour was established depending on whether the hindered train has to halt to a standstill before the red signal. Finally, absolute prediction error and its standard deviation for different prediction horizons was presented. In future work, more accurate predictions can be made by focusing precise modelling of dwell times. However, this requires measurements and data from other sources than train describers.

References

[Bue12] T. BUEKER and B. SEYBOLD: “Stochastic modelling of delay propagation in large networks”. In: *Journal of Rail Transport Planning & Management* 2.1-2 (2012), pp. 34–50.

[Dol09] U. DOLDER, M. KRISTA, and M. VOELCKER: “RCS – Rail Control System – Realtime train run simulation and conflict detection on a net wide scale based on updated train positions”. In: *Proceedings of the 3rd International Seminar*

on Railway Operations Modelling and Analyisis (RailZurich2009). Zurich, 2009, pp. 1–15.

[Gov11] R. M. P. GOVERDE and L. MENG: "Advanced monitoring and management information of railway operations". In: *Journal of Rail Transport Planning & Management* 1.2 (2011), pp. 69–79.

[Han08] I. A. HANSEN and J. PACHL, eds.: *Railway Timetable & Traffic - Analysis, Modelling, Simulation*. Hamburg: Eurailpress, 2008.

[Han10] I. A. HANSEN, R. M. P. GOVERDE, and D. J. VAN DER MEER: "Online train delay recognition and running time prediction". In: *13th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. Madeira, 2010, pp. 1783–1788.

[Kec12] P. KECMAN and R. M. P. GOVERDE: "Process mining of train describer event data and automatic conflict identification". In: *Computers in Railways XIII, WIT Transactions on The Built Environment*. Ed. by C. A. BREBBIA, N. TOMII, and J. M. MERA. Vol. 127. Southampton: WIT Press, 2012, pp. 227–238.

[Kec13a] P. KECMAN and R. M. P. GOVERDE: "Adaptive, data-driven, online prediction of train event times". In: *16th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. The Hague, 2013.

[Kec13b] P. KECMAN and R. M. P. GOVERDE: "An online railway traffic prediction model". In: *Proceedings of the 5th International Seminar on Railway Operations Modelling and Analyisis (RailCopenhagen2013)*. Copenhagen, 2013, pp. 1–17.

[Lue09] M. LUETHI: "Improving the Efficiency of Heavily Used Railway Networks through Integrated Real-Time Rescheduling". PhD Thesis. ETH Zurich, 2009.

[Med11] G. MEDEOSSI, G. LONGO, and S. de FABRIS: "A method for using stochastic blocking times to improve timetable planning". In: *Journal of Rail Transport Planning & Management* 1.1 (2011), pp. 1–13.

[Nas04] A. NASH and D. HUERLIMANN: "Railroad simulation using OpenTrack". In: *Computers in Railways IX*. Ed. by J. ALLAN, C. A. BREBBIA, R. J. HILL, G. SCIUTTO, and S. SONE. Southampton: WIT Press, 2004, pp. 45–54.

[Rou06] P. J. ROUSSEEUW and K. DRIESSEN: "Computing LTS Regression for Large Data Sets". In: *Data Mining and Knowledge Discovery* 12.1 (2006), pp. 29–45.

[Sie06] T. SIEFER and A. RADTKE: "Evaluation of delay propagation". In: *Proceedings of 7th World Congress on Railway Research*. Montreal, 2006.

Corresponding author: Pavle Kecman, Delft University of Technology, Department of Transport & Planning, Stevinweg 1, 2628 CN Delft, Netherlands phone: +31 15 278 4914, e-mail: p.kecman@tudelft.nl

Timetable Evaluation and Optimization under Consideration of the Stochastical Influence of the Dwell Times

Anne Binder, Thomas Albrecht
Technische Universität Dresden

Abstract

An approach for timetabling of regional railway lines is presented which explicitly considers the stochasticity of the dwelling process and the properties of modern driver advisory systems for energy-efficient train running time control. A stochastic optimization approach is presented to determine the expected values of the optimization goals for different timetables. The algorithm is applied in a case study for a German regional railway line.

Keywords: Allowance time allocation, multi-criterian timetabling, stochastical dwell times

1 Motivation

Within railway systems the given timetable represents the basis upon which operation is carried out [Han08]. This paper focuses on planning of regional and suburban trains which constitute the majority of European railway traffic. In the EU-project ON-TIME [OTP13] a survey among railway infrastructure managers (IMs) in Europe was carried out. One point that was observed in the answers is that it is desirable to have these regional services planned at regular headways throughout the day with easily recallable departure times [Gov13]. These departure times have severe impact on rail operation:

1. Passenger trains must never depart before the published departure time given to the railway costumer within the passenger information system.

2. Delays and delay penalties (of trains and of passengers) are determined with respect to the times published in the timetable.

3. The capability of a timetable to absorb short term requests for additional train paths is determined to a large extent by the fixed frequent service patterns.

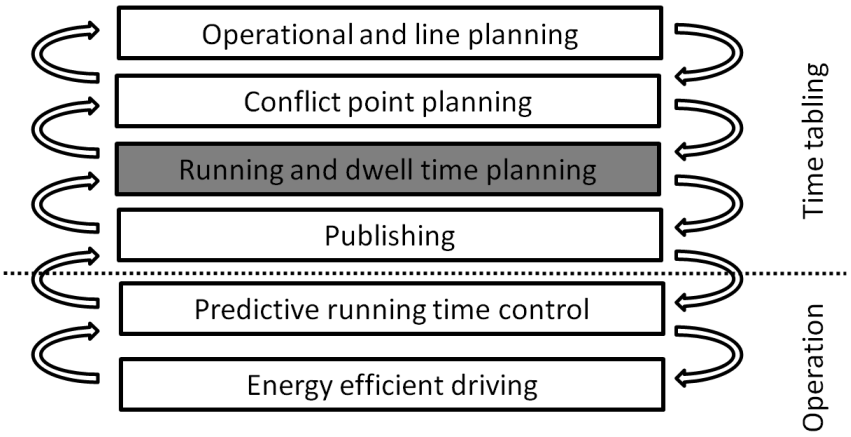


Figure 1: Hierarchy of timetabling and operation processes

The timetable and in particular the departure times are determined in a multi-stage process. First, railway undertakings (RUs) make general decisions on the lines, stops, rolling stock and train frequencies in the step called operational and line planning in Fig. 1. Then, the IM must work out a feasible timetable for all RUs operating in a network which is mainly determined by the bottlenecks in the infrastructure, where trains pass, cross or overtake. In the next two stages of finetuning of the timetable, allowances are finally allocated to the timetable and all departure times are fixed. The real-time operational control measures which make use of the allowances in the timetable are running time control and energy-efficient driving – working at the level of seconds. Both are part of a few advanced driver advisory systems for regional train operation [Bin12; Alb12].

This paper deals with the allowance allocation process or finetuning of the timetable (gray in Fig. 1). Allowances need to exist in the timetable in order to cope with statistical variations of operations. For the running times, a minimal running time can be computed, the size of running time allowances added to the minimal running time is often prescribed by some IM rules [UIC00]. For dwell times, a minimal technical duration can be computed as well. The rest of the dwell time depends on the number and distribution of alighting and boarding passengers, station and train layout and other environmental parameters which have been studied and described in the literature in multiple ways for heavy-rail or metro-like systems, e.g. [Har07]. However, for dwell time allowance allocation there are no rules known in Europe. Dwell times are often planned using standard time values per station or service kind. No experience is reported, which quantile of the dwell time distribution needs to be used for planning. Additionally, dwell time allowances can be used as running time allowances, this is particularly true for regional services and frequent stops of durations shorter than the usual precision of published timetables of 1 min. In that case, dwell and running time allowances can hardly be distinguished (from a planning point of view).

The paper presents an approach which considers the stochastic of the dwelling process already in the timetable planning phase. It starts with the description of an algorithm for

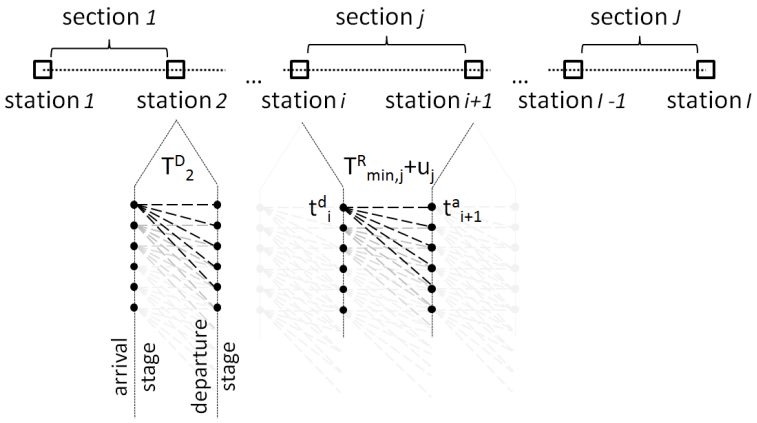


Figure 2: Definition of the search space

energy-efficient train running time allowance allocation along a line which considers dwell time uncertainties. This algorithm is constrained in its actions by the planned departure times. In section 3 it is described how an optimal timetable along a line can be obtained. The whole algorithm is applied in a case study on a regional line in Germany, the results are presented in section 4. The last section concludes the main aspects of the papers and presents ideas for further research.

2 System model

2.1 Description of the search space

Definitions

The search space represents the volume of all potential solutions of the running time optimization problem. The definitions are illustrated in Figure 2. Thereby a railway line with a total amount of I stations -each identified by i- will be considered. Station $i = I$ is defined as target station and the section between station i and $i+1$ is indexed with j. The total number of sections between station 1 and I is given by $J = I - 1$. A timetable Φ consisting of planned arrival and departure times, $t^a_{plan,i}$ and $t^d_{plan,i}$, shall be analyzed:

$$\Phi = \left\{ t^d_{plan,1} \cdots t^d_{plan,I-1}; t^a_{plan,2} \cdots t^a_{plan,I} \right\}$$

The minimal and maximal running times on the sections, $T^R_{min,j}$ and $T^R_{max,j}$, $j \in [1; J]$, and the minimal and maximal dwell times at each intermediate station, $T^D_{min,i}$ and $T^D_{max,i}$, $i \in [2; I-1]$ are given. The minimal times are defined by technical limitations, the maximal times depend on definitions given by the operator, e.g. minimum allowed speed on the sections and maximum dwell time at the intermediate stations in undisturbed operation.

Possible arrival and departure times at the intermediate stations are denominated with t^a_i and t^d_i. The total amount of allowance time U which should be allocated among the section is defined by

$$U = t^a_{plan,I} - t^d_{plan,1} - \sum_{i=2}^{I-1} T^D_{min,i} - \sum_{j=1}^{J} T^R_{min,j}. \quad (1)$$

Stage transition

In contrast to the previously published approach of the authors [Bin12] in which the dwell times are assumed as constant, the stochasticity of the dwelling process shall be considered here. Each intermediate station stop is modeled with an arrival and departure stage. The transition of the departure from a station i to the following arrival stage at station $i+1$ is assumed to be a deterministic process. This assumption can be ensured by the use of driver advisory systems or in automatic train operation. Consequently, the arrival time depends on the minimal running time and the used allowance u_j on this section j:

$$t^a_{i+1} = t^d_i + T^R_{min,j} + u_j, \quad (2)$$

$$0 \leq u_j \leq T^R_{max,j} - T^R_{min,j}.$$

Because no departure before the planned departure time shall be allowed, the departure time t^d_i does not only depend on the arrival time t^a_i and dwell time T^D_i , but is determined by

$$t^d_i = \max\left(t^a_i + T^D_i, t^d_{plan,i}\right). \quad (3)$$

In contrast to the running time, the dwell time can hardly be controlled and random deviations in the dwell time within the operation have to be handled. Consequently, the dwell time cannot be assumed to be a predictable constant value. The dwell times follow a random distribution $f_i(T^D_i)$ which reflect the variations in the dwelling process and lead to the dependency

$$t^d_i\left(t^a_i; f_i(T^D_i)\right).$$

In the system model each dwelling process is regarded as independent event in undisturbed operation below the limit of capacity. The dwell time distribution $f_i(T^D_i)$ is normalized between the defined $T^D_{min,i}$ and $T^D_{max,i}$. However, the total planned allowance time U might not be sufficient to cover all randomly occurring long dwell times along a line. There is still the probability that the trains will arrive with a delay at target station. Hence, a maximum tolerated threshold at the target station ϑ and the minimum percentage of trains Ψ which shall arrive with a maximum delay of ϑ have to be defined and limit the search space.

Limitation of the search space

The upper borders for arrival and departure times at the intermediate stations are limited by the probability $P(O_I|t^a_i)$ and $P(O_I|t^d_i)$ which illustrate the probability of the event O_I that the train will arrive at the target station I within the tolerated threshold ϑ (see formulas

(4) and (5)). If the probability is below Ψ this arrival or departure time will not satisfy the constraints given by the operator and is therefore not a possible solution. Consequently, each tolerated arrival or departure time must fulfill the equation

$$P(O_I|t_i^a) = P(t_I^a \leq t_{plan,I}^a + \vartheta) \geq \Psi, \tag{4}$$

$$P(O_I|t_i^d) = P(t_I^a \leq t_{plan,I}^a + \vartheta) \geq \Psi. \tag{5}$$

Based on the initial values for the arrival times at the target station given in (6)

$$P(O_I|t_I^a) = \begin{cases} 1 & t_I^a \leq t_{plan,I}^a + \vartheta \\ 0 & t_I^a > t_{plan,I}^a + \vartheta \end{cases} \tag{6}$$

these limits of tolerated departure stages can be calculated recursively by

$$P(O_I|t_i^d) = \begin{cases} P(O_I|t_{plan,i}^d) & t_i^d < t_{plan,i}^d \\ P(O_I|t_{i+1}^a = t_i^d + T_{min,j}^R + u_j) & t_i^d \geq t_{plan,i}^d \end{cases} \tag{7}$$

The values at the arrival stages are determined recursively, as well, by the convolution of the departure stages and the dwell time distribution at station i:

$$P(O_I|t_i^a) = \int_{T_{min,i}^D}^{T_{max,i}^D} f_i(\tau) \cdot P(O_I|t_i^d = t_i^a + \tau) d\tau. \tag{8}$$

2.2 Quality criteria determination

For quantifying the quality of a timetable, three criteria $Q_q, q \in [1, 2, 3]$ are chosen which are considered important from the operators or passengers point of view. As these criteria are influenced by the stochastic of the dwell times and the non-linear constraints of the given $t_{plan,i}^a$ and $t_{plan,i}^d$, they depend on $f_i(T_i^D)$ and cannot be determined nominally, but their expected values at each arrival and departure stage have to be calculated with a recursive approach. Thereby the constraint that early departures ($t_i^d < t_{plan,i}^d$) are not permitted has to be ensured.

Expected energy consumption

The amount of consumed energy depends on rolling stock and infrastructure characteristics of the line, but also on the available time allowance u_j on each section – which is assumed to be used for energy-efficient driving. The energy consumption $E_j(u_j)$ for a given time allowance can be calculated for each section j using algorithms described in the literature [How94].

The expected energy consumption for each arrival and departure stage can be determined by formulas (9)- (11).

$$Q^d_{1,i}(t^d_i) = \begin{cases} Q^d_{1,i}(t^d_{plan,i}) & t^d_i < t^d_{plan,i} \\ E_j(u_j) + Q^a_{1,i+1}(t^a_{i+1} = t^d_i + u_j + T^R_{min,j}) & t^d_i \geq t^d_{plan,i} \end{cases} \tag{9}$$

$$Q^a_{1,i}(t^a_i) = \int_{T^D_{min,i}}^{T^D_{max,i}} f_i(\tau) \cdot Q^d_{1,i}(t^a_i + \tau) d\tau \tag{10}$$

$$Q^a_{1,I}(\forall t^a_I) := 0 \tag{11}$$

Expected delay at target station

As described in section 2.1 delays at the target station cannot be avoided. However, an expected delay at the target station which could be important from the operators point of view can be calculated for each departure and arrival time at all intermediate stations in the following way:

$$Q^d_{2,i}(t^d_i) = \begin{cases} Q^d_{2,i}(t^d_{plan,i}) & t^d_i < t^d_{plan,i} \\ Q^a_{2,i+1}(t^a_{i+1} = t^d_i + u_j + T^R_{min,j}) & t^d_i \geq t^d_{plan,i} \end{cases} \tag{12}$$

$$Q^a_{2,i}(t^a_i) = \int_{T^D_{min,i}}^{T^D_{max,i}} f_i(\tau) \cdot Q^d_{2,i}(t^a_i + \tau) d\tau \tag{13}$$

$$Q^a_{2,I}(t^a_i) := \max(0, t^a_I - t^a_{plan,I}) \tag{14}$$

Expected delay at intermediate stations

Although the intermediate stations are regarded as minor objective concerning the operational stability, occurring delays are relevant from the passenger point of view. Therefore the total amount of expected intermediate delay is the third optimization criteria which is given by

$$Q^d_{3,i}(t^d_i) = \begin{cases} Q^d_{3,i}(t^d_{plan,i}) & t^d_i < t^d_{plan,i} \\ Q^a_{3,i+1}(t^a_{i+1} = t^d_i + u_j + T^R_{min,j}) & t^d_i \geq t^d_{plan,i} \end{cases} \tag{15}$$

$$Q^a_{3,i}(t^a_i) = \max(0, t^a_i - t^a_{plan,i}) + \int_{T^D_{min,i}}^{T^D_{max,i}} f_i(\tau) \cdot Q^d_{3,i}(t^a_i + \tau) d\tau \tag{16}$$

$$Q^a_{3,I}(\forall t^a_I) := 0 \tag{17}$$

3 Optimization Process

3.1 Multi-criteria running time allowance allocation

All quality criteria depend on the allocation of the allowance time u_j along the train run. Thereby they are modeled as "Markovian", i.e. the value of a criterion at a certain state $Q^d_{q,i}(t^d_i)$ does not depend on the previous states, but on the present state and the sequence of decisions $u_j ... u_J$ that follow. Because the all quality criteria are Markovian, the multi-stage optimization problem can be solved by dynamic programming [Bel57]. Thereby for each departure time a decision has to be made about the amount of allowance time u_j to spend until the arrival stage at the next station. The decision is made based on the expected values of the criteria within a multi-criteria approach.

Here, the weighted metrics method is used to determine a substitute objective criterion $\tilde{Q}$, as this method is supposed to be less sensitive to the choice of weighting factors w_q compared to other methods of multi-criteria decision making [Ehr05]. The optimal time allowance u^*_j is determined for each departure time by the minimization of the distance to the optimal value for each criterion:

$$\tilde{Q}(u_j, t^d_i) = \sqrt{\sum_q w_q \left(Q^d_{q,i}(u_j, t^d_i) - \min_{\forall u_j} Q^d_{q,i}(u_j, t^d_i) \right)^2} \tag{18}$$

The optimal allowance time u^*_j has to be found for each possible departure time t^d_i at each stop using the substitute optimization criterion:

$$u^*_j(t^d_i) := \min_{\forall u_j} \tilde{Q}(u_j, t^d_i) \tag{19}$$

3.2 Timetable evaluation

The output of the allowance allocation process are for each possible departure and arrival time at each station expected values of the relevant criteria for the remaining train run until the target station. Hence, the values obtained for the given departure time at the first station $Q^d_{q,1}$ give an indication about the timetable itself, because the criteria are significantly influenced by $t^a_{plan,i}$ and $t^d_{plan,i}$.

In order to find the optimal timetable, multiple timetables Φ_k can be analyzed. In this first approach, complete enumeration is used to find all potential combinations of $t^a_{plan,i}$ and $t^d_{plan,i}$ with respect to the minimal running and dwell times and following the constraint $t^a_{plan,i} \leq t^d_{plan,i}$. After the stochastic running time allocation process the optimization criteria $Q^d_{1,1}(\Phi_k)$, $Q^d_{2,1}(\Phi_k)$ and $Q^d_{3,1}(\Phi_k)$ can be used within a further multi-criteria optimization approach to determine the best timetable Φ^* (see (20)). The applied weighted metrics method requires the definition of weighting factors v_q for each timetable evaluation criteria.

$$\Phi^* := \min_{\forall \Phi_k} \sqrt{\sum_q v_q \left(Q^d_{q,1,}(\Phi_k) - \min_{\forall \Phi_k} Q^d_{q,1}(\Phi_k) \right)^2} \tag{20}$$

4 Case study on a German regional train line

The algorithm has been tested on the regional railway line from Halberstadt (LHB) to Aschersleben (LAL) in Germany which consists of six track sections. The planned travel time today is 29 minutes, the sum of minimal running times is 21 minutes with one of the diesel multiple units used here, i.e. 8 minutes must be used for dwelling at the five intermediate stops and can partly be allocated as running time allowances. The timetable is given with a precision of 1 min. The energy consumption as function of the running time has been taken from the calculations of a driver advisory system application [Alb12]. The dwell time distributions have been taken from historical data of the traffic control system. In order to obtain departure times which are good for an entire day, only one dwell time distribution per stop was used which contains data for entire operating days including morning and afternoon peak as well as the off-peak hours.

For this line, 1287 different reasonable timetables exist (combinations of planned departure times at the intermediate stops, the planned arrival times were set equal to the planned departure times). For all of them, the expected values for the optimization criteria were determined using the following parameters: $\Psi = 97\%, \vartheta = 2\ \mathrm{min}, w_1 = w_2 = w_3 = 1$ (quality criteria given by the RU). Thereby the current computation time is about 4 hours which meets the demands of the timetabling process. In Table 1 the obtained values for four timetables are given: the three timetables which represent the minima for the individual criteria and the best timetable using $v_1 = v_2 = v_3 = 1$. The departure times for each of them can be taken from Fig. 3. For each of the timetables, one specific run is illustrated in that figure, for which it was assumed that each of the dwelling processes at the intermediate stops takes as long as its 50%-quantile value.

Table 1: Expected values (Ev) and specific results (Sr) for train run with dwell time quantile 0.5 for different timetables

	Φ_1			Φ_2			Φ_3			Φ_4		
	$\mathbf{Q_1}$	Q_2	Q_3	Q_1	$\mathbf{Q_2}$	Q_3	Q_1	Q_2	$\mathbf{Q_3}$	Q_1	Q_2	Q_3
Ev	**96.8**	115.6	39.95	111.1	**0.2**	587.1	101.2	62.1	**0.2**	106.4	1.7	2.0
Sr	**95.2**	103.0	38	110.0	**0**	456	99.4	45	**0**	103.1	0	0

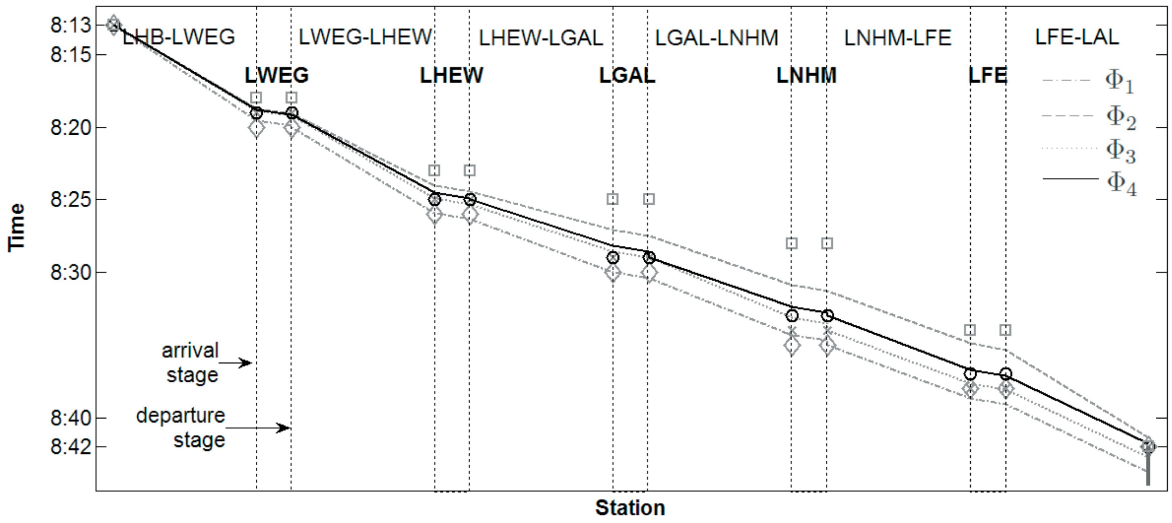


Figure 3: Schematic comparison of different timetables and corresponding optimized running times relating to dwell time quantile 0.5

4.1 Energy-optimal timetable Φ_1

The gray diamonds show the arrival and departure times for the timetable which minimizes the consumed total energy consumption during the run (dash-dotted line). As energy consumption is the single optimization criteria, the other optimization criteria are not considered besides the hard constraints (maximum tolerated delay at target station of 2 minutes is almost exploited). Therefore the target station is not reached on time, but the additional running time is used to increase the running times on almost all sections which decreases the energy consumption. The arrival and departure times at the intermediate stations are planned according to a huge running time reserve which can be used for energy efficient driving.

4.2 Timetable with minimal delay at target station Φ_2

The minimization of the expected target station delay leads to the optimal timetable which is illustrated by the square-marked arrival and departure stages and the gray dashed line. The timetable which leads to the minimal expected target station delay is a tensely planned timetable which spends all the available allowance time on the last section. In addition, as the running times are rounded down to the full minute, the planned running time is even not sufficient to be punctual at the intermediate stops. Therefore, time optimal driving is applied on every section except for the last section (LFE - LAL). No time allowance is neither used for energy efficient driving nor waiting within a station for the departure time. As an obvious result, this leads to high intermediate delay and energy consumption.

4.3 Timetable with minimal intermediate delay Φ_3

The timetable which results in minimal intermediate arrival delays is represented by the gray crosses at the stations and the dotted gray line. The calculations have shown, that for higher assumed dwell time quantiles the arrival times at the intermediate stations can be ensured, as well. Due to the fact that enough running time is allocated on each section, small intermediate departure delays at the intermediate sections (e.g. at LHEW) can be caught up until the next station. However, as for the choice of this timetable the target station delay is not regarded, the remaining running time on the last station is not enough to arrive on time at the target station.

4.4 Balanced timetable Φ_4

The resulting timetable of a balance weighting between the three criteria is illustrated by the black circles and the corresponding black time-space line. Apparently, the first intermediate arrival and departure time are similar to the timetable Φ_3. As the target station delay is considered, as well, the running time on the following sections is shortened, in order to enable a punctual arrival at the target station. In the displayed run with the 0.5-dwell time quantile the train has to wait for the planned departure time at the stations LGAL and LNHM. This leads to a higher total energy consumption, but to a significantly smaller delay at the target station compared to timetables Φ_1 and Φ_3.

5 Conclusions and outlook

The proposed algorithm considers the uncertainties of the dwelling process in the timetabling of regional railway lines explicitly and thus can provide mathematically optimized timetables. It is important to notice that opposed to todays understanding of the timetabling process, planned arrival and departure times have to be considered as goals which can be reached with certain probabilities only.

As the approach consists of two multi-criteria optimization levels, the choice of the weighting factors are a highly sensitive task. Therefore sensitive analysis towards the weighting factors will be executed. At the moment, only the dwell times are regarded stochastic and only the variations caused by the passenger boarding and alighting process have therefore been considered. In further research the approach shall be extended for stochastic running times (driving without DAS) and dwell times resulting from conflicts with other trains. Furthermore, a benchmark shall be made with other stochastic timetabling approaches where energy consumption is not considered at the moment [Vro05].

Acknowledgement

This work is part of the project ON-TIME (Optimal Networks for Train Integration Management in Europe) which is co-funded by the European Commission within the Seventh

Framework Programme (2007-2013), Grant Agreement FP7-SCP01-GA-2011-285243. The dwell time data was used with kind permission of Veolia Sachsen-Anhalt Verkehr GmbH (railway undertaking) and INTERAUTOMATION Deutschland GmbH (provider of the traffic control system).

References

[Alb12] T. ALBRECHT and M. PATHE: "Pilotierung eines Assistenzsystems zur kraftstoffsparenden Fahrweise im SPNV". In: *EI - Eisenbahningenieur* 63 (9) (2012), pp. 66–70.

[Bel57] R. BELLMAN: *Dynamic Programming*. Princeton University Press, 1957.

[Bin12] A. BINDER and T. ALBRECHT: "Predictive Energy-Efficient Running Time Control for Metro Lines". In: *CASPT12 - Conference on Advanced Systems for Public Transport*. 2012.

[Ehr05] M. EHRGOTT: *Multi-criteria Optimization*. 2. Springer-Verlag Berlin Heidelherg, 2005.

[Gov13] R. M. GOVERDE and I. A. HANSEN: "Performance Indicators for Railway Timetables". In: *IEEE International Conference on Intelligent Rail Transportation*. 2013.

[Han08] I. A. HANSEN, J. PACHL, and et AL.: *Railway timetable & traffic*. Ed. by I. A. HANSEN. 1. DVV Media Group GmbH | DVV Rail Media (Eurailpress), Hamburg, 2008.

[Har07] N. HARRIS and R. ANDERSON: "An international comparison of urban rail boarding and alighting rates". In: *Proceedings of the Institution of Mechanical Engineers. Part F, Journal of rail and rapid transit* 221(4) (2007), pp. 521–526.

[How94] P. HOWLETT, I. MILROY, and P. PUDNEY: *Energy-efficient train control*. Vol. 2. 2. 1994, pp. 193–200. URL: http://www.sciencedirect.com/science/article/pii/0967066194901988.

[OTP13] *ON-TIME Project Website*. (Last Access: 30 June 2013). URL: http://www.ontime-project.eu/.

[UIC00] *UIC Code 451 - 1: Timetable recovery margins to guarantee timekeeping - Recovery margins*. Tech. rep. UIC - Union Internationale des Chemins de fer, 2000.

[Vro05] M. VROMANS: "Reliability of Railway Systems". PhD thesis. Erasmus University Rotterdam, Erasmus Research Institute of Management (ERIM), 2005.

Corresponding author: Anne Binder, Technische Universität Dresden, "Friedrich List" Faculty of Transport and Traffic Sciences, Institute for Traffic Telematics, 01062 Dresden, Germany, phone: +49 351 463 36764, e-mail: anne.binder@tu-dresden.de

Combining Demand Management and Merge Control in an Equilibrium Network Model

Francesco Viti[1], Wei Huang[2], Mike J. Smith[3]

[1] University of Luxembourg, Luxembourg

[2] Katholieke Universiteit Leuven, Belgium

[3] University of York, United Kingdom

Abstract

Equilibrium models under congested traffic conditions, and especially those addressing blocking back, are very useful to estimate the demand conditions that ITS policies should be able to manage, for instance to maintain congestion within controlled areas and avoiding that they further spillback and cause more serious and/or less controllable congestion states.

The objective of this paper is to supplement the equilibrium model, developed by the authors in recent research, with a more thorough analysis of merge behaviour, especially in cases of blocked nodes. Regulating the merger behaviour together with the demand pattern can lead to certain desired stationary states. It has a great practical significance when congestion is inevitable, while demand management and merge control are able to retain queues and spill-backs within the local area.

Keywords: Equilibrium, Queuing, Merging, Blocking back, Merging control policy.

1 Introduction

This paper integrates and extends recent modelling developments of the authors in the area of quasi-dynamic traffic assignment problems [Smi13], which are recently being proposed as a convenient trade-off between modelling parsimony requirements sought in network equilibrium analysis and the more complex network effects caused by traffic congestion.

By adopting a novel spatial queuing approach, in our previous work we derived equilibrium conditions that explicitly consider buffer spaces occupied by queues on links, which in turn determine the extent to which vehicles move in free driving mode within a link and the capacity restrictions due to blocking back at nodes. This allowed us to study the effect of limited queue storage capacity, and to determine signal time responses to both

saturation and buffer space capacities that guarantee an equilibrium to be reached, under certain conditions.

In this study we focus on the impact of different merging policies on queue sizes and on equilibrium. Furthermore, a better understanding of the relationship between node models and equilibrium conditions allows us to assess different control strategies and give recommendations on how to manage (sub-)networks in an efficient way, either by metering competing flows so that they match both equilibrium and merging requirements, or by adopting specific signal setting policies on the merges that would guarantee solution convergence to equilibrium.

The findings in this paper are relevant for ITS as they will contribute to network-wide dynamic traffic management by means of dynamic demand and supply control strategies which will be quickly found and used to control traffic in real time. A clear advantage of these models is their simplicity and their applicability within more complex online control systems.

2 Equilibrium with queuing, block-back and capacity constraints

This section summarizes the main model developments in [Smi13], which represent the basic information used to apply and analyse different merge models.

Under congested conditions, queues emerge at bottlenecks, and under some condition they back-propagate. In line with the philosophy adopted in [Smi13] we are not considering in this study the spatio-temporal propagation of the queue fronts within a link and onto nodes. Instead, we account for the effects of space taken up by traffic queues and spillback onto nodes within an equilibrium model, which requires only the explicit calculation of steady state conditions, and not necessarily its transient evolution. This type of approach is being referred to as quasi-dynamic equilibrium formulation (see e.g. [Dag98], [Bli12]).

2.1 The basic link model

The quasi-dynamic assignment formulation developed in [Smi13] lays its foundations on a special link model, which explicitly takes into account the reduced space in free-driving mode due to the occurrence of a queue in the link.

Let v_i be the flow entering link i, s_i the saturation flow at the exit of link i, and Q_i the queue of vehicles waiting to exit link i; let the maximum possible value of Q_i be $MaxQ_i$ and the time to traverse the entire length of link i (when the queue $Q_i = 0$) is $c_i(v_i)$. Feasibility conditions are the inequalities $v_i \leq s_i$ and $Q_i \leq MaxQ_i$.

Let also b_i be the delay due to queuing experienced by vehicles, from the moment they joined the queue until they exited link i. Using the classical Little's law we can assume $b_i = Q_i/s_i$.

Consider the case where $0 \leq Q_i \leq MaxQ_i$. In this case, the queue occupies part of the length of link i, so the cost in free-driving mode tends to decrease with increasing Q_i. The

time for traversing link i will thus be the sum of a non-queueing component and a queueing component, which are mutually varying. This mutual consistency can be modelled in a rather simple manner with the following relationship, which provides the total cost spent to traverse link i:

$$Sum_i = c_i(v_i) + k_i b_i \quad (1)$$

where k_i is denoted as the 'shrinkage' factor. In [Smi13] it has been shown that in case of no blocking back this shrinkage factor takes the expression

$$k_i = 1 - s_i c_i(s_i)/MaxQ_i \quad (2)$$

Expression (2) prevents the typical overestimation problem when vertical queuing is used; in fact, if $Q_i = 0$ then $Sum_i = c_i(v_i)$, which is rather straightforward as the total costs is consisting of only the driving time to traverse the whole link as function of the flow, while if $Q_i = MaxQ_i$ then $Sum_i = MaxQ_i/s_i = Maxb_i$, so the total cost to traverse the fully congested link consists of only the queuing delay. For all values $0 < Q_i < MaxQ_i$ it can be observed that $k_i > 0$.

Note that it would be more 'natural' to adopt a shrinkage factor to multiply $c_i(v_i)$, as it seems more intuitive to observe a reduced space in free-driving mode. However, it can be shown that a multiplier $h_i = [MaxQ_i - Q_i]/MaxQ_i$ associated to $c_i(v_i)$ leads to the same expression. In addition, k_i does not depend on Q_i, which is found only solving the equilibrium problem, while expression (2) solely depends on constant and predetermined parameters. A demonstration of the equivalence between shrinkage factors k_i and h_i can be found in [Smi13].

Expressions (1)-(2) assume no blocking back, so that the condition $Q_i > 0$ can occur only if $v_i = s_i$. When link outflow is restricted by downstream queue filling a downstream link and overflowing, the flow along link i can be less than s_i. The shrinkage factor k_i becomes in this case dependent on v_i and is no longer constant:

$$Sum_i = c_i(v_i) + k_i b_i = c_i(v_i) + [1 - v_i c_i(v_i)/MaxQ_i] b_i \quad (3)$$

2.2. Network representation

The network representation in this study follows the multi-level framework used extensively in past studies to model multi-commodity flows. For more details one should look at [Smi13].

2.3. Equilibrium formulation

We use the standard Wardrop [War52] notion that for each OD pair more costly routes are unused. Again, details on equilibrium conditions and the Variational Inequality formulation derived are found in [Smi13]. For sake of keeping the paper to the required size limits we directly deal with the problem analysing a simple network example, which serves as proof of concept.

2.4 A simple numerical example

Consider the network in Figure 1; two OD pairs are joined by three routes as follows.

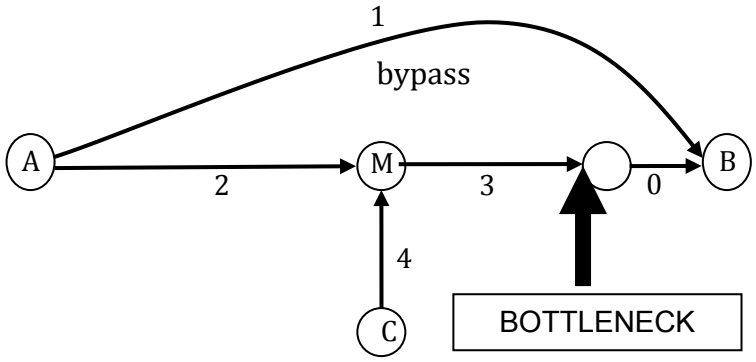


Figure 1: Simple network with one bottleneck, two origins and one destination.

- Route 1 (the bypass) joins the OD pair [A,B], and has just link 1.
- Route 2 joins OD pair [A,B], and has links 2, 3 and 0.
- Route 3 joins OD pair [C,B], and has links 4, 3 and 0.

Link 2 has a saturation flow at the exit of s_2 vehicles per minute and link 3 has a saturation flow at the exit of s_3 vehicles per minute. All other links will have very large saturation flows and $s_2 > s_3$ so the exit of link 3 is a bottleneck. Links 2 and 4 merge at M. The steady OD flow rate from A to B is fixed at T_{AB} vehicles per minute where $T_{AB} > 0$. The steady OD flow rate from C to B is fixed at T_{CB} vehicles. If v_i is the flow rate along link i, we suppose that with no queueing the time taken to traverse link i (i = 0, 1, 2, 3, 4) will be $c_i(v_i)$. We assume that link 0 has zero travel time as it does not influence the results.

In case of no merging traffic, T_{CB}=0. We assume that $T_{AB} > s_3$ so that not all the demand can be served in the considered time period (here and for all the analysis considered unitary), and that $c_2(v_2) + c_3(v_3) < c_1(v_1) \leq c_2(v_2) + Maxb_3$ so at equilibrium queuing occurs at link 3, but conditions are such that this queue does not get longer than the maximum buffer space $MaxQ_3$. To determine the bottleneck delay on link 3 at equilibrium we use link performance model (3). Then $c_1(v_1) = c_2(s_3) + c_3(s_3) + k_3 b_3$, and considering that $T_{AB} = v_1 + s_3 + Q_3$ hence

$$b_3 = \frac{c_1(T_{AB} - s_3 - Q_3) - (c_2(s_3) + c_3(s_3))}{[1 - s_3 c_3(s_3)/MaxQ_3]} \tag{4}$$

considering $k_3 = [1 - s_3 c_3(s_3)/MaxQ_3]$ by (2), and $Q_3 = b_3 s_3$ by the Little's law. Thus, the steady-state equilibrium queueing delay (and the equilibrium queue) depends on the uncongested travel times, the outflow capacity of link 3 and the storage-capacity of link 3.

We now consider the case of blocking back occurring at node M; we also assume that there is not yet merging traffic flow. Suppose that $c_2(v_2) + Maxb_3 < c_1(v_1) \leq Maxb_2 + Maxb_3$. The link 3 queue must then spillback onto link 2. The equilibrium queues will be such that the total delay incurred on the two queued links 2 and 3 equals the uncongested travel time

difference between two alternative routes joining A and B. So, now $c_1(v_1) = c_2(s_3) + k_2 b_2 + Maxb_3$ using (3). Considering now that $T_{AB} = v_1 + s_3 + Q_2 + MaxQ_3$ in this case

$$b_2 = \frac{[(c_1(T_{AB} - s_3 - Q_2 - MaxQ_3) - (c_2(s_3) + c_3(s_3))) - (MaxQ_3/s_3 - c_3(s_3))]}{[1 - s_3 c_2(s_3)/MaxQ_2]} \quad (5)$$

considering $k_2 = [1 - s_3 c_2(s_3)/MaxQ_2]$ and $Q_2 = b_2 s_3$. Queuing delay becomes this time a function of the link travel times, the link saturation flow of link 3 and the storage-capacity of both links 2 and 3. These equations show the dependence of the steady-state queue on a link upstream of the block back node M and on the queue storage-capacity of the blocked link 3. One may easily verify that equilibrium cannot be achieved in this network if $c_1(v_1) > Maxb_3 + Maxb_2$, i.e. route 1 is not appealing even with fully saturated links 2 and 3. Any queue in this case would spillback onto origin node A, and thus outside the analysed network.

We have so far analysed situations where no merging flow enters the system from node M coming from origin C. If $T_{CB} > 0$ some portion of the flow and the queue observed at link 3 is taken by this demand, and reduces the opportunity for link 2. Given therefore a fixed demand, T_{CB}, we should specify a realistic merge operation. To add this extra complexity, we need to specify a merge model. In the following section we introduce different merge models proposed in literature, and analyse how the choice of a specific model form affects the existence of equilibrium.

3 Merge models

To analyse the impact of different merge models we refer again to the example network of Figure 1, therefore we deal with the problem in this study of only two merging flows onto one capacitated link. We will elaborate a general formulation with multiple merging and diverging links in future research.

It is straightforward to observe that, to have equilibrium, and if conservation of vehicles principle holds, there must be no distinction between sending and receiving flows, as it is instead done in dynamic network loading models. No stationariety would otherwise be observed in the system and queues would be time dependent. This can be formulated as:

$$r_2 = \frac{v_2}{v_2 + v_4}, r_4 = \frac{v_4}{v_2 + v_4} \Rightarrow p = \frac{r_2}{r_4} = \frac{v_2}{v_4} \quad (6)$$

Therefore, the type of priority observed at the merge determines the proportion between flows v_2 and v_4 to be met at equilibrium. This holds for any arbitrary time period, and therefore also for the analysed unitary time period.

At equilibrium, four possible states can occur, involving two merging links: 1) No queue is observed at any of the entering links as demand is low; 2) Only link 2 is queued; 3) Only link 4 is queued; and 4) Both links are queued.

Various merge ratios are suitable for different layouts, different junction geometries and different “controls”. A variety of factors in real life determine how these output flows and

delays are distributed among the two approaching traffic streams. Here we distinguish and discuss some basic rule:

- ***Fixed merge models***

Merge ratios can be assumed constant, and the ratio p can be fixed according to observed behaviour in real merging situations. Examples of these models in literature are the Daganzo's fixed merge model [Dag98], where p can take any arbitrary value. An instance of Daganzo's fixed merge model is the "zipper" rule, which assumes that drivers give way to the other approach in such a way that vehicles from the two approaches alternate with equal frequency. The model is obtained by letting the distribution fractions in the previous model be ½.

- ***"Fair shares" merge models***

These models assume that merge behaviour is in dependent on the traffic load of each merging link, or has some relationship with the importance of the roads, normally represented by their capacity. In this class of models distribution fractions are proportional to the flows along the incoming links, i.e. the higher flow tends to get more priority. This means, in mathematical terms, that $p = v_1/(v_1 + v_2)$, or they can be proportional to the saturation flows of the incoming links, i.e. more capacitated links get more priority. This means, in mathematical terms, that $p = s_1/(s_1 + s_2)$.

- ***Equal delay merge models***

A third and perhaps more sophisticated merge rule is that merge priority is in some way proportional to the queue lengths or the delay incurred at each merging link. This rule mimics the natural behaviour of drivers who tend to get more risk prone and less inclined to give way to if they had to wait longer for their turn. In mathematical terms, this means that a new equilibrium condition is imposed in the system. In this paper we do not distinguish the two equilibrium conditions (equal queue length or equal delay) as they are equivalent due to the adoption of Little's law.

3.1 Equilibrium and merging constraints

We assume that link 1 is used and link 3 is fully saturated. We assume initially that no queues are observed at links 2 and 4, but only, eventually, at link 3. It holds straightforwardly:

$$c_1(v_1) = c_2(v_2) + c_3(s_3) + k_3 b_3 \tag{7}$$
$$T_{AB} = v_1 + r_2(s_3 + Q_3) = v_1 + \frac{p}{p+1}(s_3 + Q_3)$$
$$T_{CB} = r_4(s_3 + Q_3) = \frac{1}{p+1}(s_3 + Q_3)$$

Equations (7) are necessary conditions to observe equilibrium on the simple network in Figure 1, which satisfies the merging fraction p introduced in eqn (6) while no queue emerges on both merging links. Considering $k_3 = [1 - s_3 c_3(s_3)/MaxQ_3]$ by (2), and

$Q_3 = b_3 s_3$ by the Little's law, we obtain

$$b_3 = \frac{c_1(T_{AB} - \frac{p}{p+1}(s_3 + Q_3)) - (c_2(s_3) + c_3(s_3))}{1 - s_3 c_3(s_3)/MaxQ_3} \tag{8}$$

If instead maximum queue is reached at link 3, $MaxQ_3$, while link 2 is the only one queued (point **C**), using eqn (2) to account for the relation between free-driving and queuing delay onto the link, we obtain:

$$\begin{gathered} c_1(v_1) = c_2(v_2) + k_2 b_2 + Max b_3 \\ T_{AB} = v_1 + Q_2 + r_2(s_3 + MaxQ_3) = v_1 + Q_2 + \frac{p}{p+1}(s_3 + MaxQ_3) \\ T_{CB} = r_4(s_3 + MaxQ_3) = \frac{1}{p+1}(s_3 + MaxQ_3) \end{gathered} \tag{9}$$

resulting in

$$b_2 = \frac{c_1(T_{AB} - Q_2 - \frac{p}{p+1}(s_3 + MaxQ_3)) - (c_2(s_3) + c_3(s_3))}{1 - s_3 c_3(s_3)/MaxQ_3} \tag{10}$$

The case of a queuing delay observed only at link 4 results in

$$\begin{gathered} c_1(v_1) = c_2(v_2) + Max b_3 \\ T_{AB} = v_1 + v_2 + r_2(s_3 + MaxQ_3) = v_1 + \frac{p}{p+1}(s_3 + MaxQ_3) \\ T_{CB} = Q_4 + r_4(s_3 + MaxQ_3) = Q_4 + \frac{1}{p+1}(s_3 + MaxQ_3) \end{gathered} \tag{11}$$

Finally if at both merging links stationary queues are observed it holds:

$$\begin{gathered} c_1(v_1) = c_2(v_2) + k_2 b_2 + Max b_3 \\ T_{AB} = v_1 + Q_2 + r_2(s_3 + MaxQ_3) = v_1 + Q_2 + \frac{p}{p+1}(s_3 + MaxQ_3) \\ T_{CB} = Q_4 + r_4(s_3 + MaxQ_3) = Q_4 + \frac{1}{p+1}(s_3 + MaxQ_3) \end{gathered} \tag{12}$$

resulting in a similar formulation for b_2 as (10).

We want to stress out that the above components are summed considering a unitary time dimension. For example looking at eqn (15), one should read for the demand equations of T_{AB} and T_{CB} that they are sum of an amount of vehicles flowing ($v_1 + s_3$ and s_3, respectively), and an amount holding in queue ($Q_2 + \frac{p}{p+1} MaxQ_3$ and $Q_4 + \frac{1}{p+1} MaxQ_3$, respectively) during a certain unitary period.

Looking at eqns (7)-(12), a desirable condition in managing (sub-)networks such as the one depicted in Figure 1, would be to coordinate the inflow into such systems and the merging priorities in such a way that equilibrium could be met. This could be a very basic area traffic control strategy, which may prevent congestion to back-propagate outside of the controlled area, so that the "damage" of blocking back could be contained to a maximum acceptable extent. It could be even more desirable if an automatic and local control policy would be able to adapt the priority p parameter such that a range of feasible demand conditions could be handled and equilibrated. In the following section we aim at deriving a very simple analytical model with this goal.

3.2 A combined demand management and merge control strategy

Taking again the example network of Figure 1, a combined inflow and merge control can be designed regarding specific management objectives. For example, if higher priority is given to link 2 in order to make sure that queues stay within downstream links, the merge distribution fraction is determined to guarantee higher merge fractions to flow 2.

Here we analyse the case where we do not put priority to any of the two merging flows, and we derive analytically the conditions for the inflows and the priority p in case of fixed merge models. Considering only the case of a queue blocking node M and back-propagating onto links 2 and 4 (thus equilibrium conditions (12)), if we assume desired queue states $Q_2 \leq MaxQ_2$ to avoid spillback onto node A, and no restriction is imposed on queue Q_4 we have to simply add the constraint

$$T_{AB} - v_1 - \frac{p}{p+1}(s_3 + MaxQ_3) \leq MaxQ_2 \tag{13}$$

which sets a specific range of possible priority fractions p, given T_{AB}. Vice-versa, T_{AB} could be limited in such a way that a certain merge priority p is allowed.

If the equal delay merge model replaces the fixed merge as the merge constraint, an additional constraint determines the extent to which Q_4 can vary. Considering the concurrent use of link 3 determined by the assumed merge priority p we obtain

$$\frac{Q_2}{r_2 s_3} = \frac{Q_4}{r_4 s_3} \tag{14}$$

which, considering that $r_2 = p/(1+p)$ and $r_4 = 1/(1+p)$ it makes the simple constraint $Q_2 = pQ_4$.

More generally, from the manager's point of view, it is more desirable to control the merge behaviour in a way that by adjusting the merge distribution fraction, equilibrium results can be obtained with conditions on feasible demand sets. The P_0 control policy of [Smi79] is analysed here. The main motivation to use this classical local control policy is to complement the spillback-avoiding strategy, guaranteed by inequality (13), with a control policy aimed at maximising the total network throughput, instead of being fair towards each merging link, as guaranteed by the equal delay condition (14).

A P_0-like control policy aimed at managing the system while keeping congestion within the controlled system should take into account the different pressure coming from the merging links (represented by the link saturation flows), but in the same time guarantee that a maximum number of vehicles is sent to the downstream link, which means in Figure 1, to make sure that $s_3 + MaxQ_3$ is sent. This is achieved by adding the following extra constraint to the network equilibrium condition:

$$s_2.[r_2(s_3 + MaxQ_3)/r_2 s_3] + Q_2 = s_4.[r_4(s_3 + MaxQ_3)/r_4 s_3] + Q_4 \tag{15}$$

where the first component of each side controls the flowing part of the system, while the second depends on the queued part.

In future papers we will discuss the properties of this control strategy (especially in terms of stability) and test it onto different networks, and we will compare it with other policies such as the equal delay policy.

4 Conclusions

This paper has extended an equilibrium model for congested networks, previously developed by the authors, by analysing the impact of different merge models.

Adding merge priorities imposes additional constraints to the existence of an equilibrium. Inversely, using these constraints in combination with Wardrop conditions enables one to identify desirable control states, for which if queues emerge and eventually back-propagate onto the nodes internal to the controllable network, they are likely to stabilize and stay within the controlled area.

Future steps will be to make a more thorough analysis of uniqueness and stability of these control strategies, and to test different management objectives, which could integrate the introduced basic constraints, for instance throughput maximization.

References

[Bli12] M. C. J. BLIEMER, L. BREDERODE, L. J. J. WISMANS, and E. S. SMITS: "Quasi-dynamic network loading: adding queuing and spillback to static traffic assignment". In: *Proceedings of the 91st TRB Annual Meeting*. Washington DC, US, Jan. 22–26, 2012.

[Dag98] C. F. DAGANZO: "Queue spillovers in transportation networks with route choice". In: *Transportation Science* 32.1 (1998), pp. 3–11.

[Smi79] M. J. SMITH: "A local traffic control policy which automatically maximises the overall travel capacity of an urban road network". In: *Proceedings of the International Symposium on Traffic Control Systems.* Berkeley: University of California, 2A (1979), pp. 11–32. – and In: *Traffic Engineering and Control* 21 (1980), pp. 298–302.

[Smi13] M. J. SMITH, W. HUANG, and F. VITI: "Equilibrium in capacitated network models with queueing delays, queue-storage, blocking back and control". In: *Procedia – Social and behavioural Sciences: 20th International Symposium on Transportation and Traffic Theory (ISTTT2013).* July 2013.

[War52] J. G. WARDROP: "Some theoretical aspects of road traffic research". In: *Proceedings of the Institute of Civil Engineers Part II* 1 (1952), pp. 325–378.

Corresponding author: Francesco Viti, University of Luxembourg, Faculty of Science, Technology and Communication, L-1358 Luxembourg, Luxembourg, phone: +352 4666 44 5352, e-mail: francesco.viti@uni.lu

Conflict Areas for Macroscopic Models in Dynamic Traffic Assignment

Daniele Tiddi, Bojan Kostic, Guido Gentile
DICEA, Sapienza University of Rome

Abstract

Intersections are the most critical elements of road networks, especially in urban contexts. In the representation of junctions for macroscopic DTA models, the usual assumption is that every conflict among intersecting traffic streams of different manoeuvres is fully solved by traffic signals. This simplifies the simulation of intersections, given that only merging and diversions are to be reproduced, and most DTA models are capable of addressing these basic topologies. However, quite often in practice we have intersecting manoeuvres that comply with some precedence and/or gap-acceptance rule, even in signalized junctions. These phenomena are very well tackled by micro-simulation models, while limited research has been produced to successfully and efficiently represent nodes with conflict areas in macroscopic models for Dynamic Traffic Assignment. This article addresses the above issue, providing a new formulation for conflicting traffic streams in the context of the General Link Transmission Model. To this end, the merging model with priorities is extended by associating a capacity to each manoeuvre, while the scarce resource to be split becomes the time of the conflict area. A specific parameter to reproduce different driver behaviour from polite to aggressive is also introduced. The model has been implemented in the software for Dynamic Traffic Assignment called TRE. Numerical results are presented to show how the model works for different combinations of flows, capacities and priorities.

Keywords: junctions with conflicting manoeuvres, Link Transmission Model, Dynamic Network Loading, turn priorities, polite vs. aggressive driver behavior.

1 Introduction

Dynamic Traffic Assignment (DTA) has recently received a considerable attention as an effective tool for real-time traffic management [Gen11]. Reliable DTA models are necessary for realistic estimation of current traffic states and prediction of traffic conditions in the near future.

The node model plays a crucial role in macroscopic DTA, since most delays actually occur at intersections, especially in urban networks. A conflict area model is meant to simulate in

this framework a junction point where multiple flow streams intersect each other without signal regulation, with enough realism to correctly reproduce travel times and turn flows. A typical junction has three different types of conflicts: merging, diversion and crossing. In this paper we concentrate on crossing conflicts.

In most traffic assignment models, intersections are topological elements with no space dimension, i.e. no time or cost is spent by the users to cross them; at most, a turn delay is considered. Due to the common assumption of arc cost separability, the reciprocal influence among crossing flows is neglected, even if they cross the same junction at the same time. Some models consider a total capacity for intersection nodes, with the unrealistic assumption that the total volume crossing the node contributes to its impedance and the impedance is the same for all flows. In real traffic, vehicles interact only at specific conflict points, usually between two maneuvers, that are delayed increasingly to the opposite flow volume.

In the Highway Capacity Manual (HCM), which is often used by traffic engineers to evaluate the level of service of an intersection, a left turn adjustment factor is introduced depending on the level of protection and on the effective opposing flow. In microscopic models, a gap-acceptance approach is typically implemented, as is the case of Vissim [PTV10]. This is based on the critical gap value, which represents the minimum average headway between vehicles of the opposite stream that will be accepted by drivers to cross the conflict point.

Macroscopic models avoid to reproduce interactions among individual vehicles by adopting a representation of traffic as a mono-dimensional partly compressible fluid, to gain simulation robustness and computation runtimes. A recent approach to improve the simulations of conflict points in DTA is based on representing junctions as mini-networks [Cor12] and [Tid12]. This spatial intersection model introduces dummy nodes and links to simulate conflict areas with internal constraints. Our model further develops this approach.

The proposed conflict area model, preliminary validated against a microsimulator (Vissim), has proved to be successful in reproducing with suitable accuracy several different situations, including: non-controlled junctions, junctions with different kinds of precedence or yield-of way, such as: two-way stop junctions, two-way yield junctions, and four-way stop junctions. It can also be used with roundabouts and junctions equipped with traffic lights. Given this flexibility, it was also implemented in the node model of the General Link Transmission Model (GLTM), which is the propagation engine of the software TRE - Traffic Realtime Equilibrium, by SISTeMA.

The paper is organised as follows. The second chapter provides the mathematical formulation of the model. It also shows how conflict areas can be coded in Visum – the travel demand modelling software by PTV Group. The third chapter demonstrates the validation of the model through numerous examples and different scenarios. Last chapter provides concluding remarks.

2 Model Description

The model aims at representing the effects of different factors influencing each turning movement, and that thus affect the efficiency of the junction. It should be mentioned that there are no restrictions in the number of conflict areas in a junction, or in the number of conflict areas encountered (in a specific order) by one turn.

2.1 Mathematical Formulation

The notation used to formulate the model is presented in the Table 1 below.

Table 1: Notation and description of used terms

Notation	Description
N	the set of nodes
$i, j \in N$	generic nodes
$A \in N \times N$	the set of arcs
$a, b \in A$	generic arcs
$a[+] = a^+ \in N$	the final node, or head, of arc $a \in A$
$a[-] = a^- \in N$	the initial node, or tail, of arc $a \in A$
$G = (N, A)$	graph, representing the road network
$Y \in A \times A$	the set of turns, representing the permitted manoeuvres
$Y_i = Y[i] \subseteq Y$	the set of turns of node $i \in N$
$y \in Y$	generic turn
$i[+] = i^+ \subseteq A$	the forward star of node $i \in N$, $i^+ = \{a \in A : a^- = i\}$
$i[-] = i^- \subseteq A$	the backward star of node $i \in N$, $i^- = \{a \in A : a^+ = i\}$
$y[+] = y^+ \in A$	the final arc, or head, of turn $y \in Y$
$y[-] = y^- \in A$	the initial arc, or tail, of turn $y \in Y$
C	the set of conflict areas
C_i	the set of conflict areas of node $i \in N$
$c \in C$	generic conflict area
Y_c	the set of turns of conflict area $c \in C$
Y_c^{rem}	the remaining turns of conflict area $c \in C$
κ_y	the capacity of turn $y \in Y$
ρ_y	the capacity reduction factor of turn $y \in Y$
q_y^{send}	the total sending flow of turn $y \in Y$
q_y^{recv}	the total receiving flow of turn $y \in Y$
q_y^{rem}	the remaining sending flow of turn $y \in Y$
q_y	the actual flow of turn $y \in Y$
π_y	the priority factor of turn $y \in Y$
$\alpha_c \in [0,1]$	the driving behaviour parameter of conflict area $c \in C$

Notation	Description
ρ_c	the capacity reduction factor of conflict area $c \in C$
θ_c^{rem}	the remaining time share of conflict area $c \in C$

Note that most of these variables are temporal profiles, i.e. functions of times. However, the conflict area model is a particular specification of the node model in the Link Transmission Model, which is as usual solved for each instant (or time interval). For this reason there is no need to make explicit reference to the current instant.

The proposed model is an iterative process where the remaining demand is possibly assigned to the remaining supply. All variables with *rem* as a superscript refer to the current status of the iterative process.

The model introduces a reduction factor for the capacity of each turn; these can be asymmetrical, due to different prudential approach to the junction, decrease in speed, safety considerations, and type of control. The model takes also into account the general loss of efficiency in the usage of the junction due to the presence of the conflict area; this is provided through a reduction factor that is applied to the time share of the conflict area itself (that is the resource to split among manoeuvres).

Then, turn priority factors are included, in a way that each turn gets proportional share of remaining time of the conflict area (also depending on turn capacities).

Finally, vehicle behaviour is modelled through a coefficient that represents the "politeness" vs the "aggressiveness" of the drivers regarding the possibility of occupying the conflict area, even when there is no available space for them downstream due to queue spillback (represented in our case by a limited receiving flow). This results in blocking the conflict area with a wasting of its available time.

The mathematical formulation of the model is presented below.

$$\theta_c^{rem} \coloneqq \rho_c$$

$$q_t^{rem} \coloneqq \alpha_c \cdot q_t^{send} + (1-\alpha_c) \cdot \min\{q_t^{send}, q_t^{recv}\}$$

$$T_c^{rem} \coloneqq T_c$$

do until $T_c^{rem} = \varnothing$ **or** $\theta_c^{rem} = 0$

for each $t \in T_c^{rem}$

$$\lambda_t = \min\left\{ q_t^{rem}, \kappa_t \cdot \rho_t \cdot \theta_c^{rem} \cdot \frac{\pi_t \cdot \kappa_t \cdot \rho_t}{\sum_{u \in T_c^{rem}} \pi_u \cdot \kappa_u \cdot \rho_u} \right\}$$

$$q_t \coloneqq q_t + \lambda_t$$

$$q_t^{rem} \coloneqq q_t^{rem} - \lambda_t$$

$$\theta_c^{rem} := \theta_c^{rem} - \frac{\lambda_t}{\kappa_t \cdot \rho_t}$$

if $q_t^{rem} = 0$ **then** $T_c^{rem} := T_c^{rem} - \{t\}$

next t

loop

$$q_t := (1-\alpha_c) \cdot \min\left\{\frac{q_t^{recv}}{q_t}, \forall t \in T_c; 1\right\} \cdot q_t + \alpha_c \cdot q_t$$

The above procedure can be synthetized in the following Conflict Area Model:

$$q_t = q_t^{CAM}\left(q_u^{send}, q_u^{recv}, \forall u \in T_c\,;\, \alpha_c, \rho_c\,;\, \kappa_u, \rho_u, \pi_u, \forall u \in T_c\right), \forall t \in T_c\,.$$

2.2 Modelling Conflict Areas in Visum

To reproduce conflict areas in TRE, they first have to be coded in a data structure. In Visum this is ensured by creating five user-defined attributes for lane turns:

- **CONFLICTAREA.** Each conflict area is denoted with a string (in the examples we used capital letters: A, B…). A conflict area is composed of exclusively two manoeuvres. Flows of lane turns intersecting in a conflict area are both tagged with that string. In case a lane turn flow crosses several conflict areas, they are reported in order of crossing and are comma separated (i.e. A, B).

- **PRIORITY.** The non-negative priority factor is a parameter of proportionality, together with turn capacities, to the time share of a conflict area that is reserved to each conflicting flow.

- **REDUCTIONFACTOR.** Capacity reduction factor due to prudential approach to the junction. It acts as a multiplier of the remaining time share of conflict area. It takes values in the range of [0, 1], meaning that 1 represents no reduction in the remaining time share, and 0 means that there is no time available.

- **CONFLICTAREAREDUCTION.** The general loss of efficiency in the usage of the junction due to the presence of the conflict area.

- **DRIVERBEHAVIOUR.** Represents the human factor that affects the efficiency of a junction in case when the sending flow of a turn is constrained by its forward star, which cause disturbances and can affect the conflicting flow. This is the parameter of a conflict area, not the turn itself. Therefore, every turn of a conflict area should have the same value of the parameter. It takes values in the range of [0, 1], where 0 represents “polite” behaviour (no effect of the constrained flow on its conflicting flow) and 1 is used for “impolite” behaviour (conflicting flow is influenced by minimum capacity ratio).

Once lane turns are created, they have to be filled in with values for all conflict areas. It is done in Junction editor/Geometry/Lane turns. If no value is entered in some of the fields, it is assumed to be 1. If turn belongs to several conflict areas and only one value is entered in the field, it is assumed value for all the conflict areas. Different values for different conflict areas are comma separated.

3 Model Analysis

Several tests have been conducted to demonstrate the behaviour of the model. The structure of the junctions used in the tests is shown in the Figure 1. The model proved to be very robust during the whole testing.

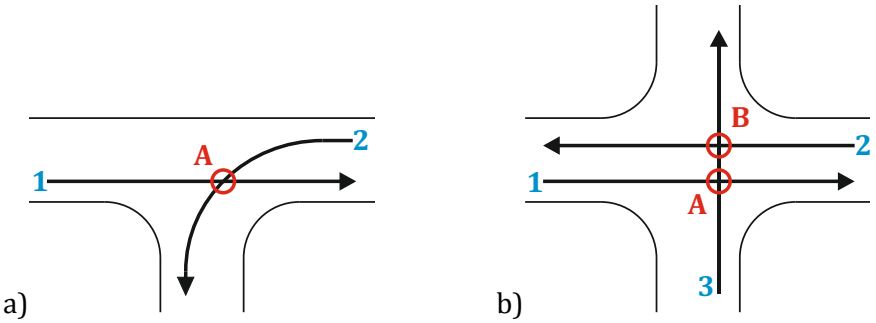


Figure 1: Test junctions: a) for scenarios 1-4 and b) for scenarios 5, 6

Scenario 1 – General conflict area effect. Here we introduce the general effect of the conflict area. All the input parameters are set to be equal. Results of this test can be used as a reference to be compared with other scenarios. Table 2 shows the values that are used in the simulation, as well as the simulation results.

Since all the supply and demand characteristics are identical, calculation for the competing flows to pass the junction is quite straightforward. From available time share of the conflict area, both flows are assigned equal share of it (50%).

Table 2: Scenario 1: input and output values

Movement	Input						Output
	κ_t	q_t^{send}	π_t	ρ_t	ρ_c	α_c	q_t
Straight move.	2000	1800	1	1	1	1	1000
Left turn	2000	1800	1	1			1000

Scenario 2 – Effect of priority (x, 0), where x > 0. The case when one priority is set to 0 and the other has positive value corresponds to the situation where the priority flow doesn't have any constraints from the conflicting flow ($q_t^{send} = q_t$). If there is any capacity left unused ($q_t^{send} < \kappa_t$), this time share of conflict area will be allotted to the non-priority flow. The table below depicts this situation. In our specific example, after the whole priority flow freely passes the intersection, the remainder of the available time share is then dedicated and used by the left turn.

Table 3: Scenario 2: input and output values

Movement	Input						Output
	κ_t	q_t^{send}	π_t	ρ_t	ρ_c	α_c	q_t
Straight move.	2000	1800	2	1	1	1	1800
Left turn	2000	1800	0	1			200

Scenario 3 – Effect of priority (x, y), where x > 0 and y > 0. When both flows have certain positive-valued priorities, calculation gets more complex. This depicts the real-world situation where non-priority flow is competing to get through the junction. This competing results in splitting the available time share of a conflict area.

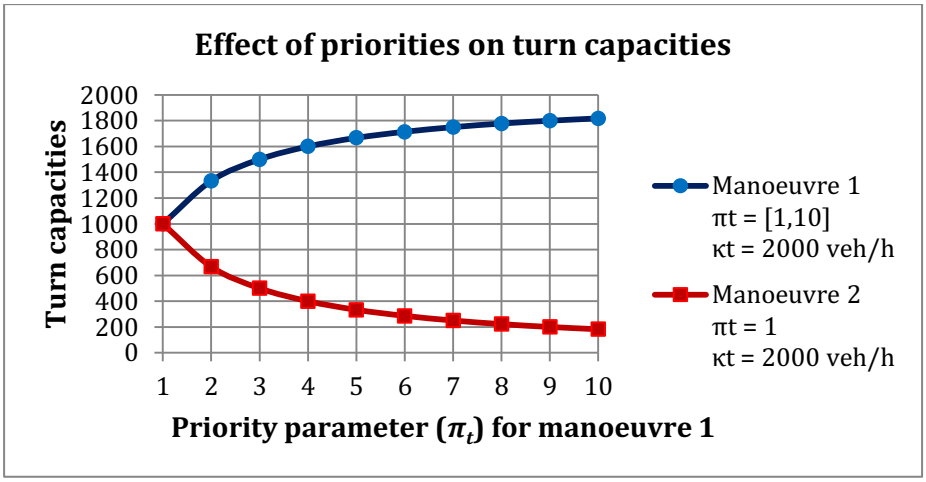


Figure 2: Scenario 3: Influence of priority parameters on turn capacities

In this analysis we show the relation between capacities in case when one turn has fixed priority parameter of 1, and the other one has increasing priority from 1 to 10. Figure 2 shows graphical results.

Scenario 4 – Effect of capacity. This represents the case where vehicles on wider lane turn have the same time available but more space to flow contemporaneously.

Table 4: Scenario 4: effect of capacity

Case	Priority parameter		Turn capacity	
	Manoeuvre 1	Manoeuvre 2	Manoeuvre 1	Manouvre 2
1	1	1	333	1333
2	2	1	500	1000
3	3	1	600	800
4	4	1	667	667
5	5	1	714	571
6	6	1	750	500
7	7	1	778	444
8	8	1	800	400
9	9	1	818	364
10	10	1	833	333

Here the turn with higher capacity (2000 veh/h) has constant priority parameter of 1. The conflicting turn has capacity of 1000 veh/h, and raise priority (Table 5).

Scenario 5 – Two conflict areas on one lane turn – the first one constraining. In this case we test the junction with two conflict areas on one lane turn. The first one is more constraining than the second one, in order to show how the non-used time share from one turn is used by the other one.

Conflict area A consists of straight turns 1 and 3, while conflict area B consists of straight turns 2 and 3 (Figure 1b). Straight turn 3 has lower priority factor in both conflict areas. For conflict area A, straight movements 1 and 3 get 1333 veh/h and 667 veh/h respectively. In this case, straight turn 2 gets 900 veh/h, while straight turn 3 gets 800 veh/h. After calculating the share of unused time (133/2000), this result is multiplied with capacity of the straight movement 2 and added to the initially calculated capacity, which ultimately gives the value of 1000 veh/h.

Table 5: Scenario 5: input and output values

Movement	Input						Output
	κ_t	q_t^{send}	π_t	ρ_t	ρ_c	α_c	q_t
Straight m. 1	2000	1800	1	1			1333
Straight m. 2	1500	1800	1	1	1	1	1000
Straight m. 3	2000	1800	0.5	1			667

Scenario 6 – Two conflict areas on one lane turn – the second one constraining. In this case, in contrast to the previous scenario, we put that the second conflict area is the constraining one, to show the spillback from one conflict area to another. Geometry and notation are the same as in the previous scenario.

Table 6: Scenario 6: input and output values

Movement	Input						Output
	κ_t	q_t^{send}	π_t	ρ_t	ρ_c	α_c	q_t
Straight m. 1	1500	1800	1	1			750
Straight m. 2	2000	1800	1	1	1	1	1333
Straight m. 3	2000	1800	0.5	1			667

Straight movement 3 has lower priority in both conflict areas. Under these circumstances, conflict area B reduces the capacity of straight movement 3, calculated for conflict area A. Going back to conflict area A, due to the minimum ratio (because of aggressive driving behaviour defined by α_c), capacity of straight movement A is reduced accordingly.

Scenario 7 – The effect of driving behaviour parameter. The effect of driving behaviour parameter (α_c) will be explained using the following example (Figure 3). Consider the case where sending flows (q_t^{send}) of both turns are set to 2000 veh/h. Receiving flow (q_t^{recv}) of constrained flow (left turn in Figure 1a) is with downstream capacity constraint (0 veh/h), and the affected flow (straight movement) has the receiving flow of 2000 veh/h. For the simplicity, all other parameters are considered to be equal. The figure below depicts how

the flow values, initial remaining flows q_t^{rem} and actual turn flows q_t, change for different values of α_c. The red line shows that the value of affected flow changes in the non-linear fashion. Further development of the model should go in the direction of ensuring linear dependence between driver behaviour parameter and actual flow of the affected turn (black line).

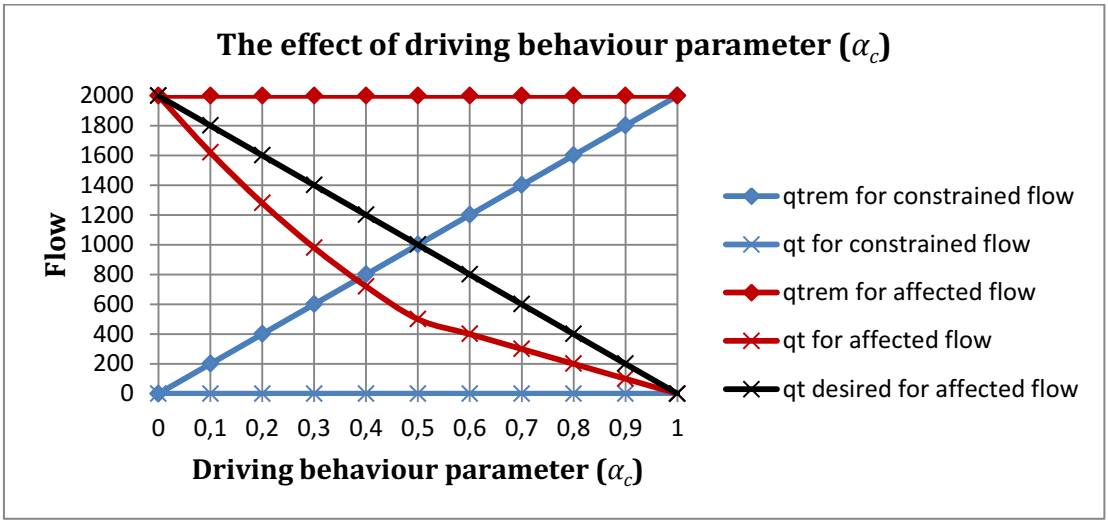


Figure 3: The effect of driving behaviour parameter (α_c)

4 Conclusion

We presented the new Conflict Area Model for modelling and simulation of conflict areas at junctions. Detailed mathematical model is shown. We introduced several factors that affect the performance of the junctions and turns. They are capacity reduction factor, due to the prudential approach to the junction; reduction factor for conflict area due to the general loss of effectiveness; priority factor, to address the precedence at the junction; and driving behaviour characteristic, which represents the "politeness" of drivers with respect to blocking the conflicting flow. Various scenarios are tested and numerical results confirmed the model to be very robust. Further developments should include corrected effect of the driving behaviour parameter.

References

[Cor12] R. CORTHOUT: "Intersection modelling and marginal simulation in macroscopic dynamic network loading". PhD thesis. Leuven, Belgium: University of Leuven, 2012.

[Gen08] G. GENTILE: "The General Link Transmission Model for dynamic network loading and a comparison with the DUE algorithm". In: *2nd International Symposium on Dynamic Traffic Assignment – DTA 2008*. Leuven, Belgium, 2008.

[Gen11] G. GENTILE and L. MESCHINI: "Using dynamic assignment models for real-time traffic forecast on large urban". In: *2nd International Conference on Models and Technolo-*

gies for Intelligent Transportation Systems. Leuven, Belgium, Jun. 22–24, 2011.

[PTV10] PTV AG: *VISSIM 5.30 User Manual*. Karlsruhe, Germany: PTV AG, 2010.

[PTV12] PTV AG: *VISUM 12.5 User Manual*. Karlsruhe, Germany: PTV AG, 2012.

[Tid12] D. TIDDI: "Models for dynamic network loading and algorithms for traffic signal synchronization in congested networks". PhD thesis. Rome, Italy: Sapienza University of Rome, 2012.

Corresponding author: Bojan Kostic, Sapienza University of Rome, Department of Civil, Constructional and Environmental Engineering (DICEA), Via Eudossiana 18, 00184 Rome, Italy, phone: +39 06 445 85737, e-mail: Bojan.Kostic@UniRoma1.it

Equilibrium and Day-to-Day Stability in Traffic Networks under ATIS

Gennaro N. Bifulco[1], Giulio E. Cantarella[2], Fulvio Simonelli[3], Pietro Velonà[4]

[1] Università degli Studi di Napoli Federico II

[2] Università degli Studi di Salerno

[3] Università degli Studi del Sannio

[4] Università degli Studi Mediterranea di Reggio Calabria

Abstract

The proposed paper describes how Advanced Traveller Information Systems (ATIS) can affect the equilibrium properties of traffic networks. The analysis is carried out within a day-to-day dynamic framework, where the equilibrium is a fixed-point state of a deterministic process, expressed as a time-discrete non-linear Markovian dynamic system. Under ATIS, original theoretical conditions are proposed for: existence and uniqueness of fixed-point states; dissipativeness of the deterministic process; stability properties and stability region of the fixed points. Identification of the stability region as a function of ATIS market penetration is of great importance from both a theoretical and practical point of view.

Keywords: ITS, ATIS, Stability Analyses, Deterministic Processes

1 Introduction

This paper develops some considerations about the topic of ATIS (Advanced Traveller Information Systems) treated in some previous works ([Bif05]; [Bif13]). Here the attention is focused on a rigorous theoretical approach to identification of some mathematical properties for both traffic equilibrium and stability. The paper collects in a coherent formulation previous findings in order to enhance the theoretical insights of some and give theoretical background and robustness to others. One of the results is identification of the stability region of the fixed points as a function of ATIS market penetration. This confirms that ATIS are a powerful tool for enhancing the stability of traffic systems, ensuring the stability of optimised network configurations which are otherwise unstable, a matter of crucial importance when traffic systems are planned and designed.

The big picture in which our work is contextualised is that of unified equilibrium and day-to-day dynamic process models, established by [Can95] and [Wat03]. Recurrent conditions are considered, meaning that the boundary conditions of the traffic system are constant, but

the system still evolves over a day-to-day dynamic process. The traffic system is considered in static conditions with respect to the within-day dynamics; the reasons for this are introduced and discussed in [Bif13]. Here we only recall that very often technicians and decision-makers plan and design transport systems under recurrent-conditions hypotheses and that within-day-static systems allow mathematical formalisations that are easier to manage from a theoretical point of view; otherwise, many of the analyses carried out here would be only based on numerical simulations.

The practical effect of the theoretical results described in this paper is that network-design options (signal-setting and/or set of one-way rules), selected by adopting an optimisation process, are actually feasible even if the equilibrium is not stable in the absence of ATIS. This means that downgrading to sub-optimal options is no longer necessary.

2 Notation and mathematical formulation

The notation here is voluntarily repeated from [Bif13], with some minor adaptation. Assume that:

i is a generic demand class with given characteristics (e.g. the O/D pair);

t is a generic simulation day in the day-to-day dynamic process;

$\boldsymbol{B}_i$ is the time-independent link-route incidence matrix for all routes (set K_i) of class i, with entries $B_{a,k}$ equal to one if link a belongs to route $k \in K_i$, zero otherwise;

d_i is the time-independent travel demand (e.g. vehic/hour) for class i ($d_i \geq 0 \ \forall \ i$);

$p^t{}_i$ is the vector of route-choice probabilities for class i at day t, with entries $p_k \ \forall \ k \in K_i$; these are computed from route travel times by using a time-independent route choice map $p_i(\cdot)$, just to fix the idea we could refer to route-choice maps based on the random utility theory (e.g.: [Dom75] or [Ben85]);

$\boldsymbol{f}^t$ is the link flows vector at day t; it belongs to the feasibility set S_f, accounting for non-negativity and ensuring both demand conservation between O/D pairs and flow conservation at nodes;

$\mathbf{c}(\cdot)$ is the congestion model, which depends on (time-independent) network-design variables (say, a vector $\boldsymbol{g}$) and gives link travel times at day t as a function of link flows at day t; it should be written as $\mathbf{c}(\cdot/\boldsymbol{g})$; however, as it is considered below that $\boldsymbol{g}$ is fixed, we omit the dependency on it;

$\mathbf{x}^t$ are the expected utilities associated to network links at day t;

$\mathbf{f}_{NL}(\cdot)$ is the network loading function; it maps link costs to the link flows ($\boldsymbol{f}$).

We refer to the simple but effective exponential-smoothing approach. It has a mathematical structure particularly suited to our theoretical goals. The dynamics of the utilities-learning process and of the choice updating process can be described (in the absence of ATIS) as:

$$\boldsymbol{x}^t = \beta \, \mathbf{c}(\boldsymbol{f}^{t-1}) + (1 - \beta) \, \boldsymbol{x}^{t-1} \qquad \text{with} \quad \boldsymbol{x}^{t=0} = \boldsymbol{x}^0 \tag{1.a}$$

$$\boldsymbol{f}^t = \alpha \, \mathbf{f}_{NL}(\boldsymbol{x}^t) + (1 - \alpha) \, \boldsymbol{f}^{t-1} \qquad \text{with} \quad \boldsymbol{f}^{t=0} = \boldsymbol{f}^0 \tag{1.b}$$

where β is the utilities-learning dynamic parameter; α is the choice-updating dynamic parameter and $\boldsymbol{x}^0$ and $\boldsymbol{f}^0$ are some known initial points of the dynamic process. In turn, the network loading function, considering the within-day stationary hypothesis, is expressed by a linear model:

$$\mathbf{f}_{NL}(\mathbf{x}^t) = \Sigma_i \, d_i \, \boldsymbol{B}_i \, \mathbf{p}_i(\, \boldsymbol{B}_i^T \, \boldsymbol{x}^t) \qquad (2)$$

In the case of ATIS some other elements have to be introduced and travelers should be distinguished with respect to their access and use of information. Three traveller groups are considered in our model. *Non-equipped travellers* do not have access to information; the fraction of these travellers is $1 - \eta$, η being the time-independent ATIS market penetration. *Compliant travellers* are equipped and make decisions whilst taking the information dispatched by ATIS into account; the fraction of these travellers at day t for class i is here identified as m^t_i ($\leq \eta$). *Non-compliant travellers* make decision without taking into account information dispatched by ATIS; the fraction of these travellers at day t for class i is $\eta - m^t_i$. Fractions of compliant travellers for different classes can be arranged, at any day t, in a vector $\boldsymbol{m}^t = [...,m^t_{,i},...]^T$. A descriptive ATIS is assumed to be in-place and the dispatched information is the ATIS-estimated travel times. The information dispatched to all equipped travellers (compliant and non-compliant) can be arranged in a vector $\boldsymbol{r}^t = [... \mid \boldsymbol{r}^t_i{}^T \mid ...]^T$, where $\boldsymbol{r}^t_i{}^T$ is the vector of information dispatched to class i (say, the *i-th* O/D pair). On the basis of the received information, compliant travellers choose their route according to a route-choice map that is $\boldsymbol{\pi}_i(\cdot)$; it is possibly different from $\mathbf{p}_i(\cdot)$ for compliant travellers; for instance, in the case of random utility, $\boldsymbol{\pi}_i(\cdot)$ could be more deterministic than $\mathbf{p}_i(\cdot)$. Under ATIS, the network loading function $\mathbf{f}_{NL}(\cdot)$ should be specialised for compliant and non-compliant travellers (or non-equipped, in the following they will not be re-specified and will be implicitly associated to non-compliant travellers), respectively $\mathbf{f}_C(\cdot)$ and $\mathbf{f}_U(\cdot)$. As in our framework compliant travellers trust the information dispatched by the system and use it for their route choice model, the (within-day static) network loading functions for compliant and non-compliant travellers are:

$$\mathbf{f}_C(\boldsymbol{r}^t, \boldsymbol{\lambda}^t) = \Sigma_i \, m_i{}^t \, d_i \, \boldsymbol{B}_i \, \boldsymbol{\pi}_i(\mathbf{r}^t_i) \qquad (3)$$

$$\mathbf{f}_U(\mathbf{x}^t, \boldsymbol{\lambda}^t) = \Sigma_i \, (1 - m_i{}^t) \, d_i \, \boldsymbol{B}_i \, \mathbf{p}_i(\boldsymbol{B}_i^T \, \mathbf{x}^t) \qquad (4)$$

Note that the network loading function for non-compliant travellers $\mathbf{f}_U(\cdot)$ is here assumed to be different from the network loading function in the absence of ATIS only because of the term $(1 - m_i{}^t)$ (compare equation 4 with equation 2 above). Route-choice maps for non-equipped and non-compliant travellers are considered as being the same, equal to $\mathbf{p}_i(\cdot)$, and invariant in the presence or absence of ATIS. This assumption was made here in order to simplify the notation. With equations 3) and 4) above, the choice to update equation 1.b can be rewritten as:

$$\boldsymbol{f}^t = \alpha \, (\mathbf{f}_C(\boldsymbol{r}^t, \boldsymbol{\lambda}^t) + \mathbf{f}_U(\boldsymbol{x}^t, \boldsymbol{\lambda}^t)) + (1 - \alpha) \, \boldsymbol{f}^{t-1} \qquad \text{with} \qquad \boldsymbol{f}^{t=0} = \boldsymbol{f}^0 \qquad (5.b)$$

A dynamic process can also be applied to the compliance, using exponential-smoothing with parameter μ and initial point $m_i{}^0$ ($\forall \, i$):

$$m_i{}^t = \mu \, \mathrm{m}(In_i{}^{t-1}, \eta) + (1 - \mu) \, m_i{}^{t-1} \qquad \text{with} \qquad m_i{}^{t=0} = m_i{}^0 \qquad \forall \, i \qquad (5.c)$$

In equation 5.c) a *compliance-function* $\mathrm{m}(In_i{}^{t-1}, \eta)$ is adopted. Compliance is considered, for

any class i, to be dependent, amongst others, on the (in)accuracy of the information system, as perceived by the travellers. The inaccuracy can be computed as:

$$In_i^t = ||\boldsymbol{r}_i^t - \boldsymbol{B}_i^T \mathbf{c}(\boldsymbol{f}^t)|| \,/\, ||\boldsymbol{B}_i^T \mathbf{c}(\boldsymbol{f}^t)|| \quad (5.d)$$

where ||.|| is the Euclidean Norm of a vector and In_i^t represents the distance between the travel times dispatched by the ATIS and the actual travel times of the network. For instance, a linearly decreasing function can be adopted, starting from the market penetration in the case of perfectly accurate information $\boldsymbol{r}_i^t = \boldsymbol{B}_i^T \mathbf{c}(\boldsymbol{f}^t)$, and reaching a null value in the case of a given *critical* (minimal) inaccuracy In_{cr}:

$$m(In_i^t, \eta) = \max(\ 0, \eta/In_{cr}^t\,(In_{cr}^t - In_i^t)\ ) \quad (6)$$

Of course, equation 6) represents an analytical approximation for much more complex phenomena; see for instance [Bif07] or [Ben12]. Renumbering equation 1.a) for the sake of consistency, the dynamic process can be summarised as:

$$\boldsymbol{x}^t = \beta\, \mathbf{c}(\boldsymbol{f}^{t-1}) + (1-\beta)\, \boldsymbol{x}^{t-1} \quad \text{with} \quad \boldsymbol{x}^{t=0} = \boldsymbol{x}^0 \quad (5.a)$$

$$\boldsymbol{f}^t = \alpha\, (\mathbf{f}_C(\boldsymbol{r}^t, \boldsymbol{\lambda}^t) + \mathbf{f}_U(\boldsymbol{x}^t, \boldsymbol{\lambda}^t)) + (1-\alpha)\, \boldsymbol{f}^{t-1} \quad \text{with} \quad \boldsymbol{f}^{t=0} = \boldsymbol{f}^0 \quad (5.b)$$

$$m_i^t = \mu\, m(||\boldsymbol{r}_i^{t-1} - \boldsymbol{B}_i^T \mathbf{c}(\boldsymbol{f}^{t-1})|| \,/\, ||\boldsymbol{B}_i^T \mathbf{c}(\boldsymbol{f}^{t-1})||, \eta) + (1-\mu)\, m_i^{t-1} \quad \text{with} \quad m_i^{t=0} = m_i^0 \quad \forall\, i \quad (5.c)$$

The information strategy determines the information ($\boldsymbol{r}^t$) dispatched by the ATIS. Several strategies can be considered. For example, exogenous a-priori-known and fixed information can be dispatched ($\boldsymbol{r}^t = \boldsymbol{r}\ \forall\ t$). This is the kind of information supplied by the majority of (static) route navigators today available on the market. As an alternative, an exogenously pre-defined day-to-day profile could be tested ($\boldsymbol{r}^t = \hat{\boldsymbol{r}}^t$), even if it seems to be of little significance for practical applications. A more practical implementation could consist in applying a smoothing filter to the dispatched information as well, with a given dynamic parameter ρ, in order to adjust the information to the observed travel times in previous days $\forall\, i\ \boldsymbol{r}_i^t = \rho\, \boldsymbol{B}_i^T c^{t-1} + (1-\rho)\, \boldsymbol{r}_i^{t-1}$. Finally, a very particular strategy is the *fully-accurate* one, where the inaccuracy is null and the compliance attains the level of market penetration $\forall\, i, t\ \ \boldsymbol{r}_i^t = \boldsymbol{B}_i^T \mathbf{c}(\boldsymbol{f}^t)$, $In_i^t = 0$, $m_i^t = \eta$.

A particular point of the dynamic process is a fixed point, where the information is fixed over time $r^t = r^{t-1} = r$, $m_i^{t-1} = m_i^t = m_i^*\ \ \forall\ t$ and $\boldsymbol{f}^{t-1} = \boldsymbol{f}^t = \boldsymbol{f}^*\ \ \forall\ t$. This corresponds to an equilibrium point:

$$\boldsymbol{f}^* = \Sigma_i\, m_i^*\, d_i\, \boldsymbol{B}_i\, \boldsymbol{\pi}_i(\mathbf{r}_i) + \Sigma_i\, (1 - m_i^*)\, d_i\, \boldsymbol{B}_i\, \mathbf{p}_i(\boldsymbol{B}_i^T \mathbf{c}(\boldsymbol{f}^*)) \quad (7)$$

With $m_i^* = m(||\boldsymbol{r}_i - \boldsymbol{B}_i^T \mathbf{c}(\boldsymbol{f}^*)|| \,/\, ||\boldsymbol{B}_i^T \mathbf{c}(\boldsymbol{f}^*)||, \eta)$

In the case of fully accurate information, a specific dynamic process can be defined:

$$\boldsymbol{x}^t = \beta\, \mathbf{c}(\boldsymbol{f}_A^{t-1}) + (1-\beta)\, \boldsymbol{x}^{t-1} \quad \text{with} \quad \boldsymbol{x}^{t=0} = \boldsymbol{x}^0 \quad (8.a)$$

$$\boldsymbol{f}_A^t = \alpha\, (\mathbf{f}_C(\boldsymbol{B}_i^T \mathbf{c}(\boldsymbol{f}_A^t), \eta) + \mathbf{f}_U(\boldsymbol{x}^t, \eta)) + (1-\alpha)\, \boldsymbol{f}_A^{t-1} \quad \text{with} \quad \boldsymbol{f}_A^{t=0} = \boldsymbol{f}_A^0 \quad (8.b)$$

It is worth noting that in equation 8.b at each day a fixed point problem has to be solved since the flow vector at day t ($\boldsymbol{f}_A^t$) is both the value and in the argument of equation 8.b, which is not the case for $\boldsymbol{f}^t$ in equation 5.b. This is due to the fact that the full accuracy hypothesis introduces the so-called anticipatory-route-guidance problem ([Bot03]) in which the dispatched information has to be consistent with the travel times it induces via the compliance and congestion mechanisms. Fixed point 7) can be rewritten in the case of fully

accurate ATIS as:

$$\boldsymbol{f}_A^* = \eta \, \Sigma_i \, d_i \, \boldsymbol{B}_i \, \boldsymbol{\pi}_i(\boldsymbol{B}_i^T \, \mathbf{c}(\boldsymbol{f}_A^*)) + (1 - \eta) \, \Sigma_i \, d_i \, \boldsymbol{B}_i \, \mathbf{p}_i(\boldsymbol{B}_i^T \, \mathbf{c}(\boldsymbol{f}_A^*)) \quad (9)$$

The model is easier if the route choice maps for compliant and non-compliant travellers coincide, $\boldsymbol{\pi}_i(\cdot) = \boldsymbol{p}_i(\cdot) \forall \, i$, and the information is still fully accurate:

$$\boldsymbol{f}_{AS}^* = (\eta + (1 - \eta)) \, \Sigma_i \, d_i \, \boldsymbol{B}_i \, \boldsymbol{p}_i(\boldsymbol{B}_i^T \, \mathbf{c}(\boldsymbol{f}_{AS}^*)) = \Sigma_i \, d_i \, \boldsymbol{B}_i \, \boldsymbol{p}_i(\boldsymbol{B}_i^T \, \mathbf{c}(\boldsymbol{f}_{AS}^*)) \quad (10)$$

Actually, previous equation 10) describes a problem in which the ATIS do not play a role with respect to the equilibrium. It exactly coincides with a traditional fixed-point problem without ATIS. Indeed this result is expected. If the route choice mechanism for informed and non-informed travellers is the same, if the ATIS dispatch the exact travel times that can be experienced on the network and if the network is in equilibrium, then there is no reason for the ATIS to play a role. The same does not hold with respect to the dynamics:

$$\boldsymbol{x}^t = \beta \, \mathbf{c}(\boldsymbol{f}_{AS}^{t-1}) + (1 - \beta) \, \boldsymbol{x}^{t-1} \quad \text{with} \quad \boldsymbol{x}^{t=0} = \boldsymbol{x}^0 \quad (11.a)$$

$$\begin{aligned} \boldsymbol{f}_{AS}^t \; &= \alpha \, (\mathbf{f}_U(\boldsymbol{B}_i^T \, \mathbf{c}(\boldsymbol{f}_{AS}^t), \eta) + \mathbf{f}_U(\boldsymbol{x}^t, \eta)) + (1 - \alpha) \, \boldsymbol{f}_{AS}^{t-1} = \\ &= \alpha \, \Sigma_i \, d_i \, \boldsymbol{B}_i \, (\, \eta \, \mathbf{p}_i(\boldsymbol{B}_i^T \, \mathbf{c}(\boldsymbol{f}_{AS}^t)) + (1 - \eta) \, \mathbf{p}_i(\boldsymbol{B}_i^T \, \boldsymbol{x}^t) \,) + (1 - \alpha) \, \boldsymbol{f}_{AS}^{t-1} \quad (11.b) \\ &\text{with} \quad \boldsymbol{f}_{AS}^{t=0} = \boldsymbol{f}_{AS}^0 \end{aligned}$$

Given that $\mathbf{c}(\boldsymbol{f}_{AS}^t) \neq \boldsymbol{x}^t$, the ATIS have an effect. It could be said that the difference is that compliant travellers have instantaneous and perfect learning of network performance, while non-compliant travellers are subject to a dynamic learning process.

3 Theoretical properties

Theoretical properties will be shown for both the equilibrium and the dynamic process under ATIS. This will be made under two main hypotheses, which are not unusual in the framework of traffic assignment theories:

I. The congestion model c(f) is continuous and continuously differentiable with respect to arc flows, with positivedefinite Jacobian Jc(f) = Jac[c(f)];

II. The route-choice map, is based on the random-utility paradigm; it is continuous and continuously differentiable; the Jacobian matrix of the route map choice for non-compliant travellers is Jip(BiT x) = Jac[pi(BiT x)] and is symmetric and negative semi-definite ($\forall$ i); the same applies for compliant travellers with reference to Jiπ(ri) = Jac[πi(ri)], and thus the Jacobian of the loading map Jf(c) is negative semi-definite too.

Theoretical conditions for existence and uniqueness of the fixed point, as well as for the stability of the equilibrium, will be discussed for the case of accurate ATIS. In this case the mathematical burden is reduced, given that equations from 8.a) to 11.b) do not actually depend on the information vector $\boldsymbol{r}$. Moreover, for the sake of simplicity, the same route-choice model is applied to both compliant and non-compliant travellers, $\boldsymbol{\pi}_i(\cdot) = \boldsymbol{p}_i(\cdot)$.

Given that the equilibrium is defined by $\boldsymbol{f}_{AS}=\boldsymbol{\varphi}(\boldsymbol{f}_{AS})$ as in equation 10), which is a standard fixed point problem, it is well known that the domain of the function is a simplex and that the codomain is contained in the domain. Moreover, from hypotheses i) and ii) function $\boldsymbol{f}_{AS}=\boldsymbol{\varphi}(\boldsymbol{f}_{AS})$ is continuous and Brouwer's theorem ([Bro12]) can be applied; thus the equilibrium exists.

Uniqueness too is ensured under usual hypotheses of the equilibrium theory. However, a weaker-than-usual condition can also be identified. Fixed point $\boldsymbol{f}_{AS}^{*} = \Sigma_i\, d_i\, \boldsymbol{B}_i\, \boldsymbol{p}_i(\boldsymbol{B}_i^{T}\, \mathbf{c}(\boldsymbol{f}_{AS}^{*}))= \boldsymbol{f}(\mathbf{c}(\boldsymbol{f}_{AS}^{*}))$ as defined in equation 10) has at most one solution if condition 12) below holds.

$$| \mathbf{I} - \mathbf{J_f}\, \mathbf{J_c} | \neq 0 \tag{12}$$

It is worth noting that hypotheses i) and ii) state that $J_f(c)$ is negative semi-definite and $J_c(f)$ is positive definite. These conditions are standard hypotheses for uniqueness of equilibrium; they ensure that the real part of matrix $J_f(c)\, J_c(f)$ eigenvalues are not positive:

$$\mathrm{Re}[\mathrm{eigenvalue}(J_f\, J_c)] < 0 \tag{13}$$

Equation 13) a fortiori implies that $\mathrm{Re}[\mathrm{eigenvalue}(J_f\, J_c)] < 1$, which implies condition 12), hence the uniqueness. This also means that uniqueness condition 12) is weaker than standard ones.

In order to assess the stability of the equilibrium, the Jacobian of the transition function of the dynamic process has to be calculated. By introducing a function $\boldsymbol{\psi_2}(\mathbf{y}) = \mathbf{y} - \alpha\, \eta\, \mathbf{f}(\mathbf{c}\,(\mathbf{y}\,))$ in equations 8.a) and 8.b), these become:

$$\boldsymbol{x}^{t} = (1 - \beta)\, \boldsymbol{x}^{t-1} + \beta \mathbf{c}(\boldsymbol{f}_{AS}^{t-1}) \tag{8.c}$$

$$\boldsymbol{\psi_2}\,(\boldsymbol{f}_{AS}^{t}) = (1 - \eta\,)\mathbf{f}_U(\boldsymbol{x}^{t})) + (1 - \alpha)\, \boldsymbol{f}_{AS}^{t-1} \tag{8.d}$$

Considering that $\alpha\ \eta$ is no greater than one by construction, condition 13) implies $\mathrm{Re}[\mathrm{eigenvalue}(J_f\, J_c)] < 1$, which in turn implies $\mathrm{Re}[\mathrm{eigenvalue}(\alpha\, \eta\, J_f\, J_c)] < 1$ thus the Jacobian of function $\boldsymbol{\psi_2}(\mathbf{y})$ is non singular, $| \mathbf{I} - \alpha\, \eta\, \mathbf{J_f}\, \mathbf{J_c} | \neq 0$, actually $| \mathbf{I} - \alpha\, \eta\, \mathbf{J_f}\, \mathbf{J_c} | > 1$ (see below). By the global inverse function theorem this condition also ensures the existence of the inverse function $\boldsymbol{\psi_2}^{-1}$. Hence the transition equations can be written explicitly as:

$$\boldsymbol{x}^{t} = \beta\, \mathbf{c}(\boldsymbol{f}_{A}^{t-1}) + (1 - \beta)\, \boldsymbol{x}^{t-1} \tag{8.e}$$

$$\boldsymbol{f}_{AS}^{t} = \boldsymbol{\psi_2}^{-1}\, [\alpha\, (1 - \eta\,)\, \mathbf{f}_U(\beta\, \mathbf{c}(\boldsymbol{f}_{AS}^{t-1}) + (1 - \beta)\, \boldsymbol{x}^{t-1})) + (1 - \alpha)\, \boldsymbol{f}_{AS}^{t-1}] \tag{8.f}$$

Given that at equilibrium xt-1= xt= c(fASt-1) and fASt-1= fASt and considering that $\mathrm{Jac}[\psi 2 - 1(y)]y=\boldsymbol{f}^{t}_{AS} = [\, \mathrm{Jac}[\boldsymbol{\psi_2}(\mathbf{y})]_{y=\boldsymbol{f}^{t}_{AS}}\,]^{-1} = [\mathrm{I} - \alpha\, \eta\, J_f\ J_c\,]^{-1}$, the Jacobian of the transition function φ can be calculated as:

$$Jac[\varphi(.)] = \begin{bmatrix} (1-\beta)I & \beta J_c \\ \alpha(1-\eta)(1-\beta)\,[I - \alpha\,\eta\, J_f J_c]^{-1}\, J_f & [I - \alpha\,\eta\, J_f J_c]^{-1}[\alpha(1-\eta)\beta\, J_f J_c + (1-\alpha)I] \end{bmatrix} \tag{14}$$

If $|J_\varphi| = |\, \mathrm{Jac}[\varphi]\, |$ is everywhere less than 1, the dynamic process is dissipative and converges toward some attractor (possibly a fixed point). Given that the determinant of any block matrix L is equal to: $|\, L\,| = \begin{vmatrix} A & B \\ C & D \end{vmatrix} = |A|\ \left|D - C\ A^{-1}\ B\right|$.

By simple algebra, the determinant of Jac [φ(.)] can be calculated as:

$$\begin{aligned} |\, J_\varphi\,| \quad &= (1-\beta)^{n}\, |\, [\mathbf{I} - \alpha\, \eta\, \boldsymbol{J}_f \boldsymbol{J}_c\,]^{-1}\, [\alpha\, \beta\, (1 - \eta)\, \boldsymbol{J}_f \boldsymbol{J}_c + (1 - \alpha)\, \mathbf{I}\,] + \\ &\quad - (\alpha\, \beta\, (1 - \eta)\, (1 - \beta))/(\,1 - \beta)\, |\, [\mathbf{I} - \alpha\, \eta\, \boldsymbol{J}_f \boldsymbol{J}_c\,]^{-1} = \\ &= (1-\beta)^{n}\, |\ [\mathbf{I} - \alpha\, \eta\, \boldsymbol{J}_f \boldsymbol{J}_c\,]^{-1}\, (1 - \alpha)\, \mathbf{I}\,| = (1-\alpha)^{n}\, (1-\beta)^{n}\ |\, [\mathbf{I} - \alpha\, \eta\, \boldsymbol{J}_f \boldsymbol{J}_c\,]^{-1}\,| \end{aligned} \tag{15}$$

where n is the dimension of the link flows vector. Hence the condition for dissipative processes is:

$$(1 - \alpha)^n (1 - \beta)^n < | [\mathbf{I} - \alpha \eta \boldsymbol{J}_f \boldsymbol{J}_c] | \qquad (16)$$

which always holds for equation 13). Indeed, consider matrix $[\mathbf{I} - \alpha \eta \boldsymbol{J}_f \boldsymbol{J}_c]$: its eigenvalues are equal to $1 - \alpha \eta$ eigenvalue$[J_f J_c]$, and, because of equation 13), have a real part greater than 1. Thus the determinant at right-hand of inequality 16) is greater than 1 while the left-hand term is evidently no greater than 1 (by definition of α and β).

It is worth noting that condition 13) for uniqueness also implies that, as already noted above:

$$|[\mathbf{I} - \alpha \eta \boldsymbol{J}_f \boldsymbol{J}_c]| > 1 \qquad (17)$$

The above equation 17) will be useful in the stability analysis below. The stability of the dynamic process could be investigated through analysis of eigenvalues (λ) of the Jacobian of the transition function at the equilibrium point. These are the solutions of the equation:

$$| J_\varphi - \lambda I |= \begin{bmatrix} (1-\beta-\lambda)I & \beta J_c \\ \alpha(1-\eta)(1-\beta)[I-\alpha\eta J_f J_c]^{-1} J_f & [I-\alpha\eta J_f J_c]^{-1}[\alpha(1-\eta)\beta J_f J_c + (1-\alpha)I] - \lambda I \end{bmatrix} = 0$$

The previous determinant can be calculated for a block matrix as:

$$| \mathbf{J}_\varphi - \lambda \mathbf{I} | = (1 - \beta - \lambda)^n \; | [\mathbf{I} - \alpha \eta \mathbf{J}_f \mathbf{J}_c]^{-1} [\alpha (1 - \eta) \beta \mathbf{J}_f \mathbf{J}_c + (1 - \alpha)\mathbf{I} + - \alpha \beta (1 - \eta) (1 - \beta)) / (1 - \beta - \lambda) \mathbf{J}_f \mathbf{J}_c] - \lambda \mathbf{I} |$$

After some algebra:

$$| \mathbf{J}_\varphi - \lambda \mathbf{I} | = b/a$$

where $a = | [\mathbf{I} - \alpha \eta \mathbf{J}_f \mathbf{J}_c] |$ and

$b = | (1 - \alpha - \lambda) (1 - \beta - \lambda) \mathbf{I} - [\lambda \alpha \beta (1 - \eta) - (1 - \beta - \lambda) \lambda \alpha \eta] \mathbf{J}_f \mathbf{J}_c |$

Given that a is positive because of equation 17), the eigenvalues λ are the solution of the equation $b = 0$. Eigenvalues λ can be related to the eigenvalues (γ) of the matrix $\mathbf{J_f J_c}$; as for each γ_k two λs can be calculated. Indeed, consider equation 18) below:

$$(1 - \alpha - \lambda) (1 - \beta - \lambda) = [\lambda \alpha \beta (1 - \eta) - (1 - \beta - \lambda) \lambda \alpha \eta] \gamma_k \qquad (18)$$

Substituting equation 18) in b we obtain:

$$| (1 - \alpha - \lambda) (1 - \beta - \lambda) \mathbf{I} - [\lambda \alpha \beta (1 - \eta) - (1 - \beta - \lambda) \lambda \alpha \eta] \mathbf{J}_f \mathbf{J}_c | = = [\lambda \alpha \beta (1 - \eta) - (1 - \beta - \lambda) \lambda \alpha \eta]^n | \gamma_k \mathbf{I} - \mathbf{J}_f \mathbf{J}_c |$$

However, $| \gamma_k \mathbf{I} - \mathbf{J}_f \mathbf{J}_c |$ is null by definition of eigenvalue γ. Then the searched eigenvalues λs can be obtained by solving equation 18). As 18) is quadratic, for each γ, two λs can be obtained:

$$\lambda_{k1} = L_1 (\alpha, \beta, \eta, \gamma_k) \qquad \lambda_{k2} = L_2 (\alpha, \beta, \eta, \gamma_k) \qquad (19)$$

where α, β, η are real scalars, while γ_k is complex. The stability condition to be imposed is that the maximum modulus of λs is less than 1:

$$\max_k \{\max\{| \lambda_{k1}, \lambda_{k2}|\}\}<1 \qquad (20)$$

Decomposing γ_k in its real and imaginary part (γ^R_k, γ^I_k), making L_1 and L_2 explicit, condition 20 is:

$$\frac{\left[\frac{\vartheta}{\alpha\cdot\beta}\gamma_k^R-\left(1-\left(\frac{1+(1-\alpha)\cdot(1-\beta)}{\alpha\cdot\beta}\right)\right)\right]^2}{\left(\frac{1+(1-\alpha)\cdot(1-\beta)}{\alpha\cdot\beta}\right)^2}+\frac{\left[\frac{\vartheta}{\alpha\cdot\beta}\gamma_k^I\right]^2}{\left(\frac{1-(1-\alpha)\cdot(1-\beta)}{\alpha\cdot\beta}\right)^2}+\frac{\eta}{[1-(1-\alpha)^2\cdot(1-\beta)^2]^2}\cdot\psi\left(\gamma_k^R,\gamma_k^I,\alpha,\beta,\eta\right)<1 \quad (21)$$

where

$$\psi(\gamma_k^R,\gamma_k^I,\alpha,\beta,\eta)=\psi_1(\gamma_k^R,\gamma_k^I,\alpha,\beta,\eta)\cdot\psi_2(\gamma_k^R,\gamma_k^I,\alpha,\beta,\eta)+\psi_3(\gamma_k^R,\gamma_k^I,\alpha,\beta,\eta)$$

$$\psi_1(\gamma_k^R,\gamma_k^I,\alpha,\beta,\eta)=\alpha^2\cdot\eta\cdot\left(\gamma_k^{R\,2}+\gamma_k^{I\,2}\right)-2\cdot\alpha\cdot\gamma_k^R$$

$$\psi_2(\gamma_k^R,\gamma_k^I,\alpha,\beta,\eta)=\cdot-\alpha^2\cdot\eta^2\cdot\left(\gamma_k^{R\,2}+\gamma_k^{I\,2}\right)+2\cdot\alpha\cdot\eta\cdot\gamma_k^R-2\cdot[1-(1-\alpha)^2\cdot(1-\beta)^2]^2+$$
$$+\left(2-\alpha-\beta+\vartheta\cdot\gamma_k^R\right)^2+\vartheta^2\cdot\gamma_k^{I\,2}$$

$$\psi_3(\gamma_k^R,\gamma_k^I,\alpha,\beta,\eta)=\alpha\cdot(1-\alpha)\cdot(1-\beta)\cdot$$
$$\cdot\left\{2\cdot\left[\left(2-\alpha-\beta+\vartheta\cdot\gamma_k^R\right)^2-\vartheta^2\cdot\gamma_k^{I\,2}\right]\cdot\gamma_k^R+4\cdot\left(2-\alpha-\beta+\vartheta\cdot\gamma_k^R\right)\cdot\vartheta\cdot\gamma_k^{I\,2}\right\}$$

$$\vartheta=\alpha\cdot\beta\cdot(1-\beta)-\alpha\cdot\eta\cdot(1-\beta)$$

Condition 21) identifies in the Argand plane the boundary of the stability region as a *quartic*. In the absence of ATIS (η = 0), the quartic is an ellipse as expected ([Can97]). The quartic can be explored with respect to ATIS market penetration (η) once α and β have been fixed. For instance, assuming α = 0.5 and β = 0.6, four different patterns can be viewed in Figures 1.a to 1.d below for η = {0.2, 0.29, 0.32, 0.57}, where the no-ATIS ellipse is also shown in red. From figure 1.a to 1.c the stability region grows and rapidly includes the whole region with a negative real part, that is the actually interesting part, given that under conditions i) and ii) the eigenvalues γ_k cannot have a positive real part. In correspondence to η=0.57 the stability region covers almost all the Argand plane (Figure 1.d).

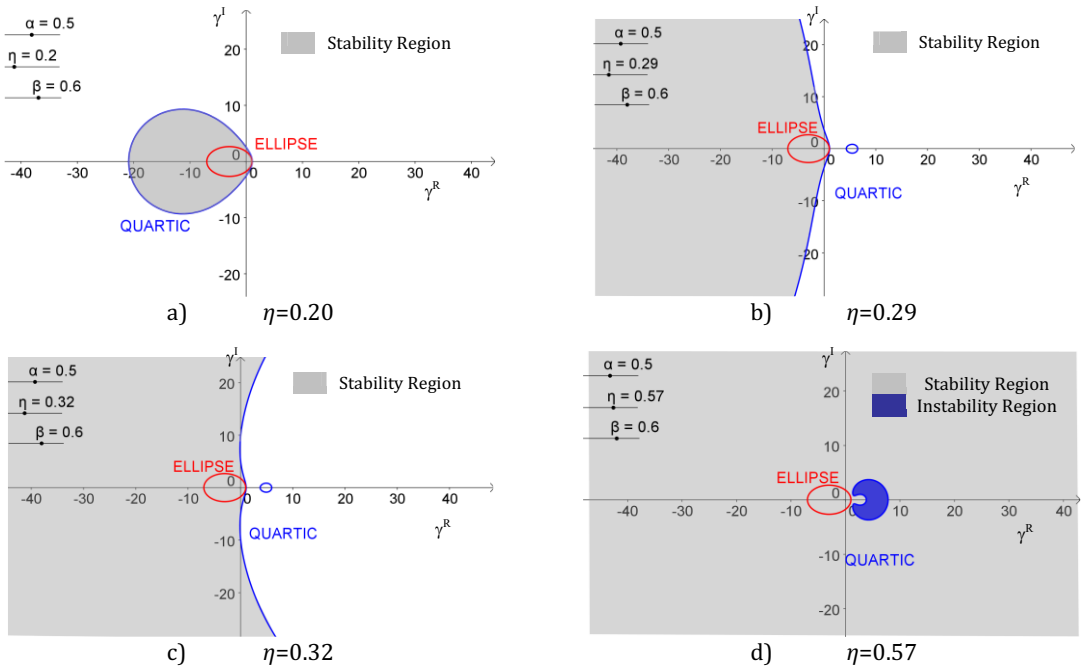


Figure 1: Stability domain depending on ATIS market penetration (α = 0.5 and β=0.6).

4 Conclusions

A dynamic process in the presence of ATIS was formalised, as well as the corresponding equilibrium model. In the particular case of accurate ATIS with the same route choice model for informed and non-informed travellers, the presence of information has no effect on the equilibrium pattern. By contrast, ATIS have a significant impact on the dynamic process and on the stability of the equilibrium. This impact can be identified in a theoretical way. The stability region can be described as a function of (amongst others) ATIS market penetration. Unlike non-ATIS networks, the stability region is a quartic (instead of an ellipse) that rapidly grows with increasing market penetration. The stability induced by ATIS can be exploited in order to stabilise network-design solutions that are, *per se*, not stable.

References

[Ben85] M. Ben-Akiva and S. R. Lerman: *Discrete choice analysis*. Cambridge, MA, USA: MIT Press, 1985.

[Bif05] G. N. Bifulco and F. Simonelli: "The effect of ATIS on transportation systems: theoretical analysis and numerical applications". In: *Urban transport XI, the built environment* 77 (2005), pp. 230–240.

[Bif07] G. N. Bifulco, F. Simonelli, and R. Di Pace: "Endogenous Driver Compliance and network performances under ATIS. Intelligent Transportation Systems Conference". In: *IEEE Intelligent Transportation Systems Conference 2007*. Vol. 33. Seattle, USA, 2007, pp. 1028–1033.

[Bif13] G. N. Bifulco, G. E. Cantarella, and F. Simonelli: "Design of signal setting and advanced traveler information systems". In: *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations.* (Accepted author version posted online: 13 May 2013). URL: http://www.tandfonline.com/doi/abs/10.1080/15472450.2013.802156#.UcVA7Pn0GCm

[Bot03] J. Bottom, S. Kachani, and G. Perakis: "A Fluid Model for the Anticipatory Route Guidance Problem". In: *10th IFAC Symposium on Control in Transportation Systems.* 2003, pp. 145–150.

[Bro12] L. E. J. Brouwer: "Ueber Abbildungen von Mannigfaltigkeiten". In: *Math. Ann.* 71 (2012), pp. 97–115.

[Can95] G. E. Cantarella and E. Cascetta: "Dynamic process and equilibrium in transportation networks: towards a unifying theory". In: *Transportation Science* 31 (1995), pp. 107–128.

[Can97] G. E. Cantarella: "Fixed-point stability and bifurcations in dynamic processes for traffic assignment". In: *Advances in Intelligent Systems*. Ed. by F. C. Morabito. Amsterdam, Netherland: IOS-press, 1997, pp. 254–261.

[Dom75] T. A. Domencich and D. McFadden: *Urban travel demand: a behavioural analysis*. Edition. Amsterdam, Netherland, 1975.

[Wat03] D. WATLING and M. L. HAZELTON: "The dynamics and equilibria of day-to-day assignment models". In: *Network and Spatial Economics* 3 (2003), pp. 349–370.

Corresponding author: Gennaro Nicola Bifulco, Università degli Studi di Napoli Federico II, Dipartimento di Ingegneria Civile Edile ed Ambientale, 80125 Napoli, Italy, phone: +39 081 7683883, e-mail: gennaro.bifulco@unina.it

Modelling Rerouting Phenomena through Dynamic Traffic Assignment in Rolling Horizon

Guido Gentile[1], Rafał Kucharski[2], Lorenzo Meschini[3]

[1] DICEA, Sapienza University of Rome

[2] Cracow University of Technology

[3] SISTeMA, PTV Group

Abstract

This article addresses the problem of simulating en-route path choices on transport networks. In particular, by rerouting, we mean changing the currently chosen path, after receiving some information about a traffic event. Indeed, when the forecasted performance pattern of travel times and costs, known or only perceived, changes significantly, drivers may react by shifting their current route to a better one. The representation of such situations is particularly challenging if the information reaches a driver who is already travelling toward the destination. At the state-of-the-practice, most traffic assignment models are not capable of reproducing these phenomena. We will model rerouting in the framework of within-day Dynamic Traffic Assignment (DTA). Two different solutions are presented here, both exploiting the rolling horizon technique. The first solution can be summarized as an alternate sequence of two fixed point problems for each traffic event: a Dynamic User Equilibrium (DUE) with warm start through saved flows, and a Dynamic Network Loading (DNL) for given route choices; this model is called DTA with Rolling Horizon Events (DTA-RHE) and allows setting the information time of each event before or after the event itself. The second one is a simplified version of the first one, under the assumption that all the events are communicated not later than their start time; this model is called DUE with Rolling Horizon Events (DUE-RHE), and is a sequence of Dynamic User Equilibria with warm start. Numerical examples show the results of the proposed models where rerouting phenomenon can be observed.

Keywords: traffic events and information, en-route path choice, Dynamic User Equilibrium, Dynamic Network Loading, Link Transmission Model.

1 Introduction

The main objective behind the proposed models is to reproduce drivers' behaviour in terms of en-route path choices when facing unexpected events in traffic networks. By unexpected event we mean any relevant traffic information that is not known in advance by at least some percentage of drivers and implies changes in the perception of the supply side, as well as different travel times. This can include: incident, road closure, longer queue, different signal plan, sport event, demonstration, planned road work, etc. In the above cases, the driver behaviour is the following: travel along the usual route, until the information of the event is received; then find a new route subject to known events, as well as their forecasted consequences, and follow it, until the next information is received, or the destination is reached.

Classic paradigms of Dynamic Traffic Assignment (DTA) are incapable of handling such cases. In a destination based dynamic Route Choice Model, the shortest trees are calculated for the given arrival time. This implies that the route is chosen at the origin based on future states of the network, taken from the previous iteration of the supply model – i.e. from previous experiences, in behavioural terms. It implicitly means that the drivers departing from origin know the state of the network up to the time of arrival at the destination. For recurrent traffic congestion this is perfectly consistent with the learning process that defines the routing behaviour on the next day, based on the experience acquired in previous days. Unfortunately such paradigm is wrong for modelling unexpected traffic events, which by definition are not known in advance nor recurrent, but change significantly the performance pattern of travel cost and times for that day.

For these cases, we propose to perform a set of simulations executed in rolling horizon, where the information on unknown events cannot be used by travellers before its communication. Our model is capable of updating route choices based on new information about the network state and apply them to traffic flows which are already on the network.

2 Dynamic Traffic Assignment

This section introduces the mathematical formulation of two versions of Dynamic Traffic Assignment (DTA) as a fixed point problems, namely the Dynamic User Equilibrium (DUE) and the Dynamic Network Loading (DNL), whose framework will be utilized in the reminder of the paper. In this paper, the focus is on the demand side, mostly Route Choice Model (RCM) and Network Flow Propagation (NFP), while the supply side, with the Link Transmission Model (LTM) is treated here as a black-box.

As the analysis is carried out within a dynamic context, all model variables are temporal profiles, here represented as piecewise C^1 functions of the time variable τ. Users trips on the road network are modelled through a strongly connected oriented graph $G = (N, A)$, where N is the set of the nodes and A is the set of the arcs.

Notation:

$q_{od}^{dem}(\tau)$ demand of users from origin o to destination d willing to depart at time τ

$p_{od}{}^{dem}(\tau)$	probability of the above users to actually make the trip; produced by the Elastic Demand Model
$q_{od}(\tau)$	flow of users travelling from origin *o* to destination *d* departing at time τ
$p_{ad}(\tau)$	probability of using arc *a* conditional on being at its initial node at time τ for users travelling to destination *d*; produced by the Route Choice Model
$q_{ad}(\tau)$	inflow of arc *a* at time τ of users travelling toward *d*
$q_a(\tau)$	inflow of arc *a* at time τ, produced by the Network Flow Propagation
$q_o(\tau)$	flow of users departing from origin *o* at time τ
$p_a(\tau)$	splitting rate of arc *a* at time τ
$\delta_a(\theta)$	characteristics of arc *a* at time θ
$f_a(\theta)$	inflow of arc *a* at time θ; produced by the Link Transmission Model
$e_a(\theta)$	outflow of arc *a* at time θ; produced by the Link Transmission Model
$t_a(\tau)$	travel time of arc *a* for users entering it at time τ
$c_a(\tau)$	cost of arc *a* for users entering it at time τ
$c_{id}(\tau)$	minimum cost to travel from node *i* to destination *d* departing at time τ
$n_{ad}(\tau)$	number of vehicles on arc *a* travelling towards destination *d* at time τ
$n_a(\tau)$	number of vehicles on arc *a* at time τ
τ_e	time at which *e*-th event is communicated
$\delta_a{}^e(\theta)$	characteristics of arc *a* expected after the *e*-th event is communicated
$c_a{}^e(\tau)$	cost of arc *a* expected after the *e*-th event is communicated
$t_a{}^e(\tau)$	travel time of arc *a* expected after the *e*-th event is communicated
$t_a{}^{real}(\tau)$	real travel time of arc *a* for users entering it at time τ

The DTA model consists of the following sub-models:

- Link Transmission Model (LTM) – can be any kind of model which takes as an input: splitting rates $p_a(\tau)$, demand flows departing at origins $q_o(\tau)$, network characteristics $\delta_a(\theta)$, and allows to obtain for each arc the performances as a function of time τ. In our case the General Link Transmission Model [Gen10b] is used to yield inflows $f_a(\theta)$ and outflows $e_a(\theta)$; we use here a different symbol for time just to emphasize that the LTM is typically implemented at a much denser time discretization than the rest of the assignment model (θ is in the order of seconds, while τ is in the order of minutes). This model can be easily substituted by a micro or meso simulation.

- First-in-first-out Travel Time (FTT) – can be seen as complementary model to calculate travel times $t_a(\tau)$ from the results of the LTM in terms of cumulative inflows and outflows [Gen05].

- Arc Cost Function (ACF) – in general works on each element of the network separately and calculates, starting from $t_a(\tau)$, its generalized cost $c_a(\tau)$ based on network parameters and user preferences. The non separability in time and space of the supply model is in the travel times.

- Route Choice Model (RCM) – is destination based and it calculates the arc conditional probabilities $p_{ad}(\tau)$ for given performance pattern of travel times and costs. Dynamic shortest trees are calculated preliminary to obtain trajectories for specific arrival times at the destination. Then a Dial like algorithm can be used to obtain a Logit loading on the efficient arcs, thus passing from deterministic to stochastic model [Bel05, Gen07]. As an alternative, a temporal layer approach can be adopted. The latter allows for a simpler presentation, but its practical implementation with discrete time intervals introduces relevant approximations (systematic flow shifts in time) with respect to the continuous theoretical solution; see [Gen04] for details.

- Network Flow Propagation (NFP) – loads the demand $q_{od}(\tau)$ towards a single destination using the arc probabilities $p_{ad}(\tau)$ from RCM and travel times $t_a(\tau)$ from LTM. When adopting a trajectory approach, the travel times from each node to the destination are actually provided by the RCM for given arrival times. NFP calculates arc inflows $q_{ad}(\tau)$ destination by destination.

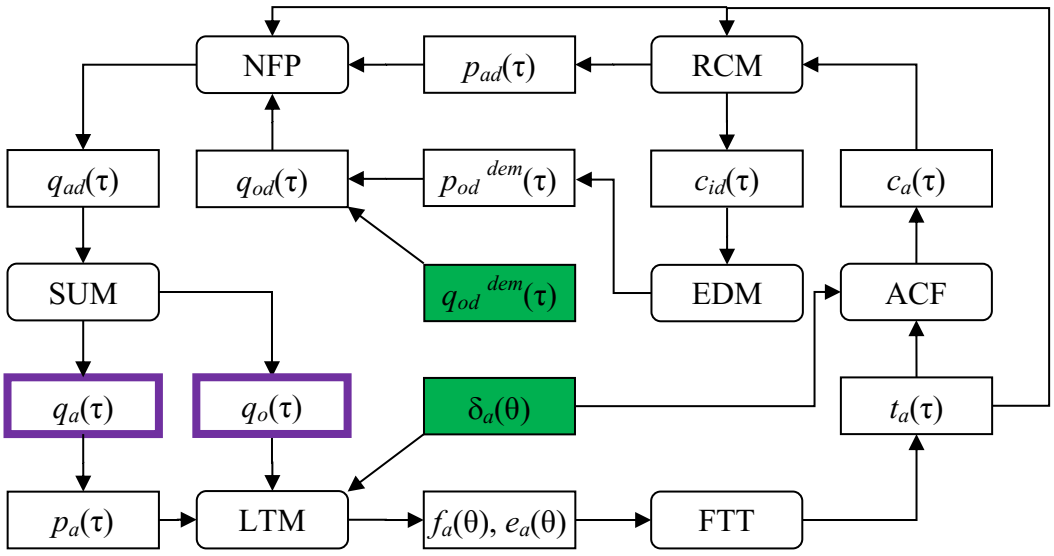


Figure 1: Fixed point formulations of Dynamic Traffic Assignment.

- Aggregation (SUM) – simply sums-up the destination specific inflows $q_{ad}(\tau)$ into origin flows $q_o(\tau)$ and arc inflows $q_a(\tau)$. The latter are used to compute the splitting rates $p_a(\tau)$.

- Elastic Demand Model (EDM) – is an optional sub-model that computes the actual travelling flows $q_{od}{}^{dem}(\tau)$ based on potential demand and travel costs, through the probability $p_{od}{}^{dem}(\tau)$ to make the trip. The latter are the result of a stochastic discrete choice model.

The Dynamic User Equilibrium (DUE) can be then formalized as a fixed-point problem in terms of the arc (and origin) flows, as shown in Figure 1:

DUE = LTM → FTT → ACF → RCM (→ EDM) → NFP → SUM → [MSA] → LTM .

The Dynamic Network Loading (DNL) is a sub-problem of DUE, which consists of seeking, for given route choices, an arc flow pattern consistent with the travel times through the arc performance model. DNL can be seen as a simplified DUE, without route-choice RCM. However, it has still a circular dependency to be solved iteratively to guarantee temporal consistency (not more than few iterations in practice). Arc (and origin) flows can be again considered as pivot variables of this fixed-point problem, as shown in Figure 1:

DNL = LTM → FTT → ACF → NFP →SUM → [MSA] → LTM .

Both fixed point problems, DUE and DNL, can be solved through the Method of Successive Averages (MSA). Although this algorithm does not converge very well in practice, it is a very flexible and robust tool; no handy alternative is yet available.

3 Rolling Horizon Assignments

A Rolling Horizon Assignment (RHA) can be seen as a sequence of fixed point DTA models. For each new (set of) information regarding traffic events in chronological order of communication, we introduce a new restart.

The first proposed solution is called Dynamic Traffic Assignment with Rolling Horizon Events (DTA-RHE). For n restarts, $2\cdot(n+1)$ classic DTA runs are needed, with saving flows for the instant τ_e of information relative to the e-th event. At each run different network characteristics need to be considered for LTM (the real travel times) and RCM (the perceived performances) to obtain appropriate results. DTA-RHE is an alternate sequence of two fixed point models for each event: a Warm DUE and a Cold DNL, as defined below.

3.1 Warm DUE

The Warm DUE (scheme of Figure 2) is a Dynamic User Equilibrium with warm start based on saved number of vehicles $n_{ad}(\tau_e)$ directed to each destination, that runs from τ_e (the time when the e-th event is communicated) to the end of the simulation. It allows to determine the performance pattern in terms of travel cost $c_a^e(\tau)$ and times $t_a^e(\tau)$ estimated by users to make their routing choices, taking into account the first e events that have been already communicated.

Thus δ_e does not include the real state of the network, but only the base scenario δ_0 and the first e known events – note that an event can be communicated after its start time. This performance pattern should also reflect the idea that user have on the consequences of communicated events. In Warm DUE we assume that each user is capable to forecast not only the direct consequences of events he gets informed of, but also the reaction of the other users. This strong hypothesis can be alleviated by performing a smaller number of DUE iterations, like if the learning process has not been completed. For $e = 0$ the Warm DUE is the base DUE.

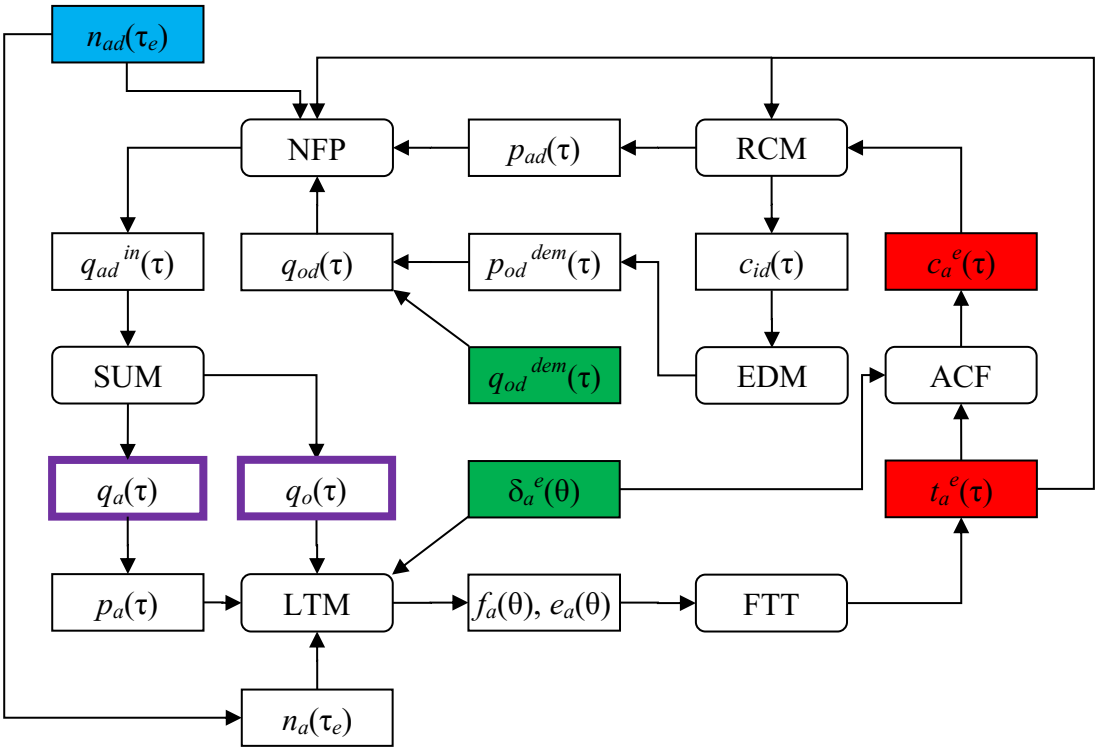


Figure 2: Warm Dynamic User Equilibrium.

3.2 Cold DNL

The Cold DNL (scheme of Figure 3) is a Dynamic Network Loading that runs from τ_e to τ_{e+1}. It allows to simulate the network and propagate the flows under real-time modifications of the supply δ^{real} with respect to the expected scenario. Modifications of arc flows due to traffic measures are also allowed in the LTM, but they will affect drivers flow patter only through travel times. The demand side (route choice and elastic trips) is instead consistent with the performance pattern $c_a^e(\tau)$, $t_a^e(\tau)$ computed in the previous Warm DUE. The Cold DNL reproduces the state that actually occurs on the network, rather than a perceived state. Travel choices are stochastic; this allows to retrieve through the RCM and the EDM the same pattern $p_{ad}(\tau)$, $p_{od}^{dem}(\tau)$ resulting from the Warm DUE for given performances $c_a^e(\tau)$, $t_a^e(\tau)$.

The idea underlying the Cold DNL is that until a user does not get a new information at time τ_{e+1} he will follow the choice pattern based on his current status of knowledge $c_a^e(\tau)$, $t_a^e(\tau)$, although the actual travel times and costs may be different from those expected. In particular, the route choice pattern is given by the arc probabilities $p_{ad}(\tau)$; that is, a user does not necessarily follow a given path but adapts his route accordingly with the times he reaches the different nodes. Travel times may actually change, given that real events occur in the meanwhile and can change the characteristics of the arcs and the flows on the network. Note that the DNL is different from the LTM since propagating the flows accordingly with $p_a(\tau)$ under the occurrence of travel times changes does not guarantee that the OD matrix is satisfied.

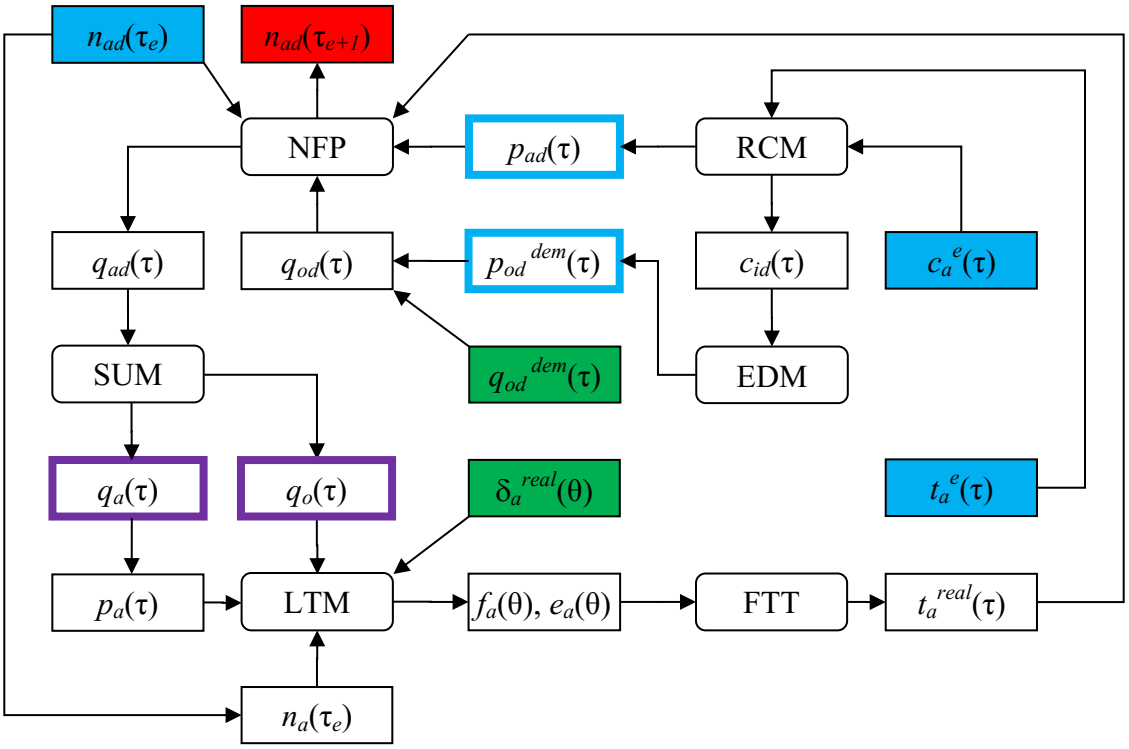


Figure 3: Cold Dynamic Network Loading.

3.3 DUE with Rolling Horizon Events

If we assume that all events are communicated not later than their start time (perfect traffic information), then there is no need to make a distinction between reality and perception (like in DTA-RHE), while the only difference with a classical DUE is that some users are informed of events after they already began their trip.

For this case we then propose an alternative model, called Dynamic User Equilibrium with Rolling Horizon Events (DUE-RHE). This is a sequence of DUE with warm start in rolling horizon, where each simulation, possibly performed in real time at time τ_{e+1} as in [Gen11], starts at time τ_e , incorporates (in particular in the LTM) all events that are communicated in the interval [τ_e , τ_{e+1}), makes a picture of the flows $n_{ad}(\tau_{e+1})$ – i.e. the number of vehicles for each arc distinguished by destination – at time τ_{e+1} , that will be the next re-starting time. The scheme of this model is depicted in Figure 4.

4 Numerical examples

Both models DTA-RHE and DUE-RHE were tested on a toy network using the software Traffic Realtime Equilibrium (TRE), by SISTeMA (www.sistemaits.com), which includes also procedures for DUE and DNL.

The toy network (see Figure 5) was designed for a single OD pair with two connections: lower - fast and efficient, and upper - alternative used only when lower connection is affected by the event. 'Escape-links' between two routes are inefficient to travel from O to D, they can be useful only to avoid event effects on lower route. The assignment lasts one hour, the event

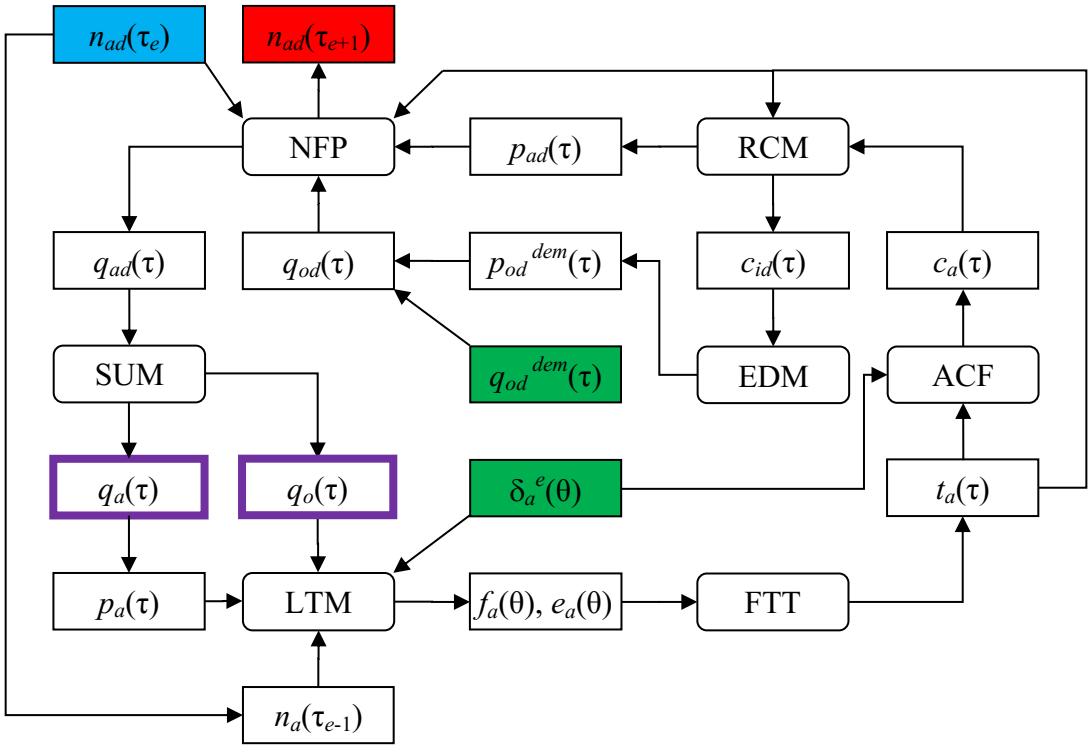


Figure 4: DUE-RHE – Dynamic User Equilibrium with Rolling Horizon Events.

happens at 40th minute at the lower route and cause speed drop to 5km/h (from 50 km/h). When the event is active, major connection is no longer efficient and the optimal choice is to a) take upper route at the origin, or b) escape lower route at the earliest convenience for users already on the network.

Two simulation results are presented below (figures 6-9): a) for an event communicated with 5 minute delay, and b) for an event known far in advance. Results for three links are presented here: 1) lower link towards the event, 2) 'escape-link', 3) initial link of upper route, thick vertical line indicates the time of the event.

If the event is known far in advance, people will reroute even before the event happens (~31st min) and no one uses 'escape-links'. In this case, the event doesn't cause any congestion, because by the time of the event the affected link is already empty. When everyone is informed, results of DTA-RHE and DUE-RHE do not vary from DUE.

On the contrary, DTA-RHE simulation of unknown event shows that users use 'escape-links' to avoid the effects of the event and reroute only after the event is communicated (45th minute). Flow is propagated accordingly with the base DUE route choice for five minutes after the event happened (during this time it was affecting supply side - LTM, but not demand side - RCM) causing congestion at the event link. At τ_e the new RCM is calculated and upper route becomes effective. The escape links are now used by users who were at the lower route at the time of event communication. Flows departing from origin after τ_e use upper route, no one chooses lower route after the event is communicated. This situation cannot be simulated by DUE-RHE or DUE.

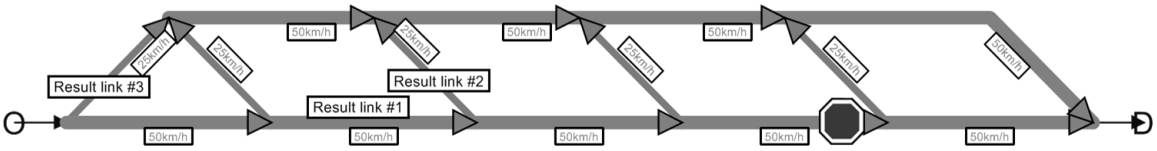


Figure 5: The toy network.

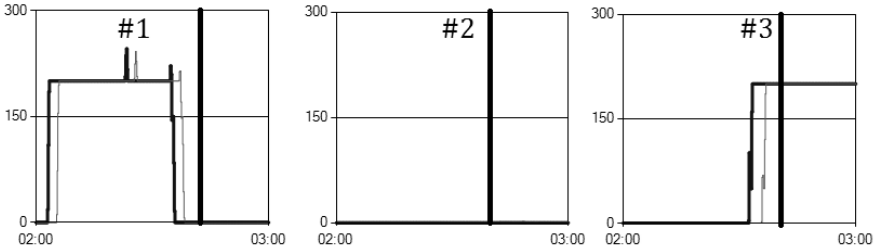


Figure 6: Flows against time at three result links for event communicated far in advance.

Figure 7: Flows on the network at the time when event communicated far in advance happens (40th minute).

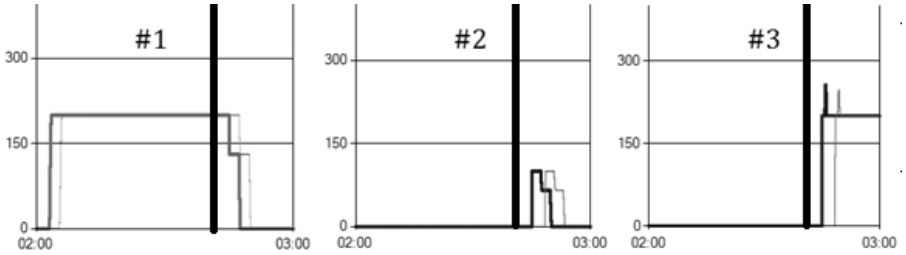


Figure 8: Flow against time at three links for event communicated with 5 minute delay.

Figure 9: Flows on the network shortly after unknown event is communicated (47th minute).

References

The short list of references presented below is only aimed at showing the research stream followed by the authors. A proper literature review goes beyond the scope of the extended abstract.

[Bel05] G. BELLEI, G.GENTILE and N. PAPOLA: "A within-day dynamic traffic assignment model for urban road networks". In: *Transportation Research Part B* 39 (2005), pp. 1-29.

[Gen04] G. GENTILE and L. MESCHINI: "Fast heuristics for continuous dynamic shortest paths and all-or-nothing assignment". In: *Proceedings of the AIRO 2004*. Lecce, Italy, 2004.

[Gen05] G. GENTILE, L. MESCHINI, and N. PAPOLA: "Macroscopic arc performance models with capacity constraints for within-day dynamic traffic assignment". In: *Transportation Research B* 39 (2005), pp. 319–338.

[Gen07] G. GENTILE, L. MESCHINI, and N. PAPOLA: "Spillback congestion in dynamic traffic assignment: a macroscopic flow model with time-varying bottlenecks". In: *Transportation Research Part B* 41(2007), pp. 1114–1138.

[Gen10a] G. GENTILE: "Traffic assignment: can dynamic models offer more?". In: *Proceedings 5th International Symposium of Networks for Mobility.* Stuttgart, Germany: FOVUS, 2010, pp. 1–5.

[Gen10b] G. GENTILE: "The General Link Transmission Model for Dynamic Network Loading and a comparison with the DUE algorithm". In: *New developments in transport planning: advances in Dynamic Traffic Assignment*, *Transport Economics, Management and Policy Series.* MA, USA: Edward Elgar Publishing, 2010, pp. 153–178.

[Gen11] G. GENTILE and L. MESCHINI: *"*Using dynamic assignment models for real-time traffic forecast on large urban networks*"*. In: *Proceedings of the 2nd International Conference on Models and Technologies for Intelligent Transportation Systems*, Leuven, Belgium, 2011.

Corresponding author: Rafał Kucharski, Cracow University of Technology, Institute of Road and Railway Engineering, Department of Transportation Systems, 30-101 Kraków, Poland, phone: +48 50 1694896, e-mail: rafalkucharski@gazeta.pl

Travel Demand for a One-Way Vehicle Sharing System: a Model of Traffic Assignment to a Multimodal Network with Supply-Demand Equilibrium

Fabien Leurent
Université Paris-Est, LVMT, Ecole des Ponts ParisTech

Abstract

A one-way Vehicle Sharing System (VSS) is strongly linked to the street network which provides not only the conditions of travel and of terminal access but also modal alternatives. Thus the system demand and especially the origin-destination pattern of flow between the VSS stations stems from the multimodal conditions and the multimodal travel demand. A three-layer framework is introduced to model a VSS in a multimodal setting, involving (i) a local station model that is a simple, Markovian system of double-ended capacitated queue [Leu12], (ii) a parking model taken from [LeB12] for customers alighting from the VSS mode with eventual cruising for parking, (iii) hyperpath choice on the multimodal network. "Availability hyperpaths" (AP) make the ex-ante travel options, from which stem the ex-post paths possibly with local adaptations due to cruising for parking [Leu13]. In an AP, local path bundling may arise from the probability of availability of a vehicle at a station for boarding. System state and traffic equilibrium are addressed in a static setting. The existence of equilibrium is demonstrated and a Method of Successive Algorithm is provided as computation method. A stylized instance is dealt with for numerical illustration.

Keywords: Vehicle sharing, Network assignment, Capacitated traffic equilibrium, Cruising for parking, Availability hyperpath

1 Introduction

To the potential user, a Vehicle Sharing System (VSS) provides transportation options to travel along some parts, say legs, in a path from origin to destination. Such paths constitute intermodal trip options, which are compared by the trip-maker with alternative options involving different combinations of Walk and perhaps other public modes, in order to select a preferred option. By using a VSS leg, the trip-maker also contributes to the service operations by picking up a vehicle at the boarding station thereby providing an available locker and by

dropping off the vehicle at the station of alighting, thereby providing an available vehicle. The quality of service of the VSS leg is assessed by the trip-maker on the basis of the probability of availability of a vehicle at the leg tail and of that of a locker at the leg head, together with the expected wait time at each leg conditional on unavailability and the generalized time along the leg. The explicit representation of resource availability constitutes a refinement over previous travel demand models of VSS within a multimodal network, e.g. [Cia13].

The paper provides an analytical framework and a simulation model for the supply and demand of VSS legs in a network of VSS stations embedded in a multimodal transportation network within an urban area. The framework involves three layers as follows: (i) the individual VSS station is modeled as a simple Markovian system, (ii) the network of VSS stations makes a system of parking options for a customer willing to alight, with diversion from a saturated station to neighboring alternatives if the expected wait time is too costly, (iii) by origin-destination pair and user class, path options are identified along the multimodal network and evaluated prior to the trip on the basis of their expected generalized cost, and every user is assigned to a path of minimum cost to him.

In a static setting, the system state is a four-fold vector that is made up of, respectively: (i) the vector of flows by link and destination on the upper layer, leg-based multimodal network, (ii) the vector of assignment proportion by destination, current node and travel strategy, (iii) the vector of assignment proportion by VSS leg and sub-network route, (iv) the vector of vehicle shortage probability by VSS station. Traffic equilibrium is defined as the joint fulfilment of all conditions within and between the sub-systems. It is characterized by a Variational Inequality Problem, of which the cost function is continuous. Then an equilibrium must exist under mild condition. A Method of Successive Averages is put forward as an Equilibration Algorithm.

The rest of the paper is organized in six sections that deal with, respectively (1) the station model, (2) the VSS mode, (3) the multimodal network, (4) traffic equilibrium, (5) a toy instance, (6) conclusion.

2 A vehicle sharing station as a dual waiting system

In [Leu12] a stochastic model of a vehicle sharing station has been designed to derive indicators of performance and quality of service from macroscopic parameters. Its core assumptions are as follows:

- The station has a docking capacity of κ docks (or lockers).
- Candidate customers arrive according to a Poisson process with time intensity λ. If no vehicle is available then a proportion r accept to wait, whereas the remaining $1-r$ choose to divert to external options.
- Leaving customers i.e. riders arrive according to a Poisson process with time intensity μ. If no dock is available then a proportion s accept to wait whereas the remaining $1-s$ divert to another station.

From these assumptions stems a bi-sided queuing system, of which the state variable is the number of busy docks i.e. of available, docked vehicles, extended to negative values for customers waiting for a vehicle and to values beyond capacity for excess vehicles waiting for a dock. The system has a stationary distribution which is unique and comprised of three pieces, each of which is a geometric (sub-)distribution:

$$p_{-k} = p_0 \rho^k \quad \forall k \geq 0 \text{, wherein } \rho \equiv \lambda r / \mu . \tag{1a}$$

$$p_n = p_0 \varphi^n \quad \forall n \in [0, \kappa] \text{, wherein } \varphi \equiv \mu / \lambda . \tag{1b}$$

$$p_{\kappa+m} = p_\kappa \sigma^m \quad \forall m \geq 0 \text{, wherein } \sigma \equiv \mu s / \lambda . \tag{1c}$$

$$p_0 = [\frac{1}{1-\rho} + \frac{\varphi - \varphi^\kappa}{1-\varphi} + \frac{\varphi^\kappa}{1-\sigma}]^{-1} . \tag{1d}$$

The outcomes of interest to us pertain to (i) the probability of vehicle shortage at the instant of customer arrival, denoted α^-, (ii) that of dock shortage, denoted α^+, (iii) the average wait time conditional on vehicle shortage, w^-, (iv) that conditional on dock shortage, w^+:

$$\alpha^- \equiv \sum_{k \geq 0} p_{-k} = p_0 / (1-\rho) . \tag{2a}$$

$$\alpha^+ \equiv \sum_{k \geq 0} p_{\kappa+k} = p_0 \varphi^\kappa / (1-\sigma) . \tag{2b}$$

$$w^- = 1/(\mu - \lambda r) . \tag{2c}$$

$$w^+ = 1/(\lambda - \mu s) . \tag{2d}$$

Furthermore, two conditions of compatibility must hold between the macroscopic parameters for the stationary distribution to exist:

$$\rho < 1 \text{ i.e. } \lambda r < \mu . \tag{3a}$$

$$\sigma < 1 \text{ i.e. } s\mu < \lambda . \tag{3b}$$

The detailed derivation of the model is included in [Leu12] together with a sensitivity analysis. The simple macroscopic properties (2) enable us to consider the station model as a sub-model in a network model. It is analogous to a travel time function for a roadway link in a static network: the network induces the link flow, from which the link sub-model (travel time function) derives the individual travel time along the link, which is the link feature relevant in the network setting.

3 Vehicle sharing system as a travel mode

VSS network. Let us assume that a one-way VSS mode is provided by a set I of stations i for access and egress, together with a network of roadway links – possibly a specific sub-network with right of way that may differ from that of the general roadway traffic.

VSS demand. To a user, the VSS mode provides travel conditions along entry-exit pairs $\ell \approx (i,s) \in \mathrm{L} \equiv \mathrm{I} \times \mathrm{I}$: these are called legs in analogy with transit services. The demand

consists in a vector $\mathbf{x}_{\mathrm{L}} = [x_\ell : \ell \in \mathrm{L}]$ of trip flows: these could be further disaggregated by demand segment according to trip destination or user class. Let us assume that along $\ell \approx (i,s)$ the customer has obtained a vehicle at i and wants to get a dock at s or in a neighboring station, in order to arrive as a pedestrian at the "landing node" s' associated to alighting from station s: the alighting link (s,s') is a pedestrian link that can be used only after dropping off the vehicle at a dock available at s. Between landing nodes n' of alternative stations $n \in \mathrm{I}$ and s', pedestrian paths are available with travel cost c'_{ns}.

Parking and routing model. The exit station s, or equivalently its landing node s', is analogous to a destination node in the Parking and Routing model of [LeB12]. The sub-set I_s of "neighboring stations" which a customer considers as attractive for leaving the vehicle and coming to s' is a "catchment area" of s as an exit station. According to [LeB12], the customer destined to s makes a two-stage choice of, first, a target station $n \in \mathrm{I}_s$ from which he begins to search for a slot and, second, a "main" path from i to n. At n he drops off the vehicle if a dock is available or he decides to wait or he chooses to divert to a neighboring station $m \in \mathrm{I}_s$ - thus making a transition $\tau \approx (n,m)$ with generalized cost g_τ.

Flows by destination station. Conditional on exit station s, let β_n^s be the probability to drop off the vehicle at n (either immediately or after some wait time w_n^-). With probability $\overline{\beta}_n^s \equiv 1 - \beta_n^s$, the customer diverts to a neighboring station; the conditional probability to choose station m, π_{nm}^s, is related to the transition cost and the expected cost from m to s' on the basis of a Discrete Choice Model.

Denote also c_n the generalized cost of drop-off at n and g_r that of route r from i to n.

The search process can be analyzed by focusing on the vector of candidate flows by exit station, $\mathbf{y}^s \equiv [y_n^s : n \in \mathrm{I}_s]$. At n the number of candidates, y_n^s, is made of the ex-ante candidates, q_n^s, plus the candidates diverted from unsuccessful requests, the n-th component of $\mathbf{y}^s \overline{\mathbf{B}}^s \mathbf{P}^s$, in which $\overline{\mathbf{B}}^s$ is the diagonal matrix of term $\overline{\beta}_n^s$ for $n \in \mathrm{I}_s$ and $\mathbf{P}^s \equiv [\pi_{nm}^s : n,m \in \mathrm{I}_s]$. Thus, in vector form, $\mathbf{y}^s = \mathbf{q}^s + \mathbf{y}^s \overline{\mathbf{B}}^s \mathbf{P}^s$ or equivalently, denoting by $\mathbf{U}^s$ the identity matrix on I_s, $\mathbf{y}^s (\mathbf{U}^s - \overline{\mathbf{B}}^s \mathbf{P}^s) = \mathbf{q}^s$. It has been shown in [LeB12] that $\mathbf{U}^s - \overline{\mathbf{B}}^s \mathbf{P}^s$ is invertible. Denoting by $\mathbf{H}^s$ its inverse matrix, it holds that

$$\mathbf{y}^s = \mathbf{q}^s \mathbf{H}^s . \tag{4}$$

From this stem the search flows of customers cruising for parking, along the transition $\tau \in \mathrm{T}_s \equiv \mathrm{I}_s \times \mathrm{I}_s$: denote by $\vec{\mathbf{P}}^s$ the matrix made up by juxtaposition of square blocks indexed by $n \in \mathrm{I}_s$, each of which is null save for its n-th row that is taken from $\mathbf{P}^s$:

$$\mathbf{x}_{\mathrm{T}}^{s} = \mathbf{y}^{s}\overline{\mathbf{B}}^{s}\vec{\mathbf{P}}^{s} = \mathbf{q}^{s}\mathbf{H}^{s}\overline{\mathbf{B}}^{s}\vec{\mathbf{P}}^{s}. \tag{5}$$

Costs by destination station. The search cost amounts to the vector product of $\mathbf{x}_{\mathrm{T}}^{s}$ times the vector of generalized transition costs, $\mathbf{c}_{\mathrm{T}}^{s}$, so

$$\widetilde{c}_{s}(\mathbf{q}^{s}, \overline{\mathbf{B}}^{s}, \mathbf{P}^{s}, \mathbf{c}_{\mathrm{T}}^{s}) = \mathbf{x}_{\mathrm{T}}^{s}.\mathbf{c}_{\mathrm{T}}^{s} = \mathbf{q}^{s}\mathbf{H}^{s}\overline{\mathbf{B}}^{s}\vec{\mathbf{P}}^{s}\mathbf{c}_{\mathrm{T}}^{s}.$$

A search that starts from station n corresponds to a particular demand vector $\delta_{n}^{s} = [1_{\{m=n\}} : m \in \mathrm{I}_{s}]$, hence to a particular search cost as follows:

$$\widetilde{c}_{n}^{s} = \widetilde{c}_{s}(\delta_{n}^{s}, \overline{\mathbf{B}}^{s}, \mathbf{P}^{s}, \mathbf{c}_{\mathrm{T}}^{s}) = \delta_{n}^{s}\mathbf{H}^{s}\overline{\mathbf{B}}^{s}\vec{\mathbf{P}}^{s}\mathbf{c}_{\mathrm{T}}^{s}. \tag{6}$$

Moreover, the "final" cost of parking due to drop off and the eventual terminal transition by walk, also depends on demand vector $\mathbf{q}^{s}$ through the derived vector $\mathbf{y}^{s}$: letting $\mathbf{B}^{s} \equiv \mathrm{diag}[\beta_{n}^{s} : n \in \mathrm{I}_{s}]$ and $\mathbf{c}_{\mathrm{I}(s)}^{s}$ be the vector of terminal costs w.r.t. alighting station,

$$\hat{c}_{s}(\mathbf{q}^{s}, \mathbf{B}^{s}, \mathbf{P}^{s}, \mathbf{c}'^{s}_{\mathrm{I}(s)}, \mathbf{c}_{\mathrm{I}(s)}^{s}) = \mathbf{y}^{s}\,\mathbf{B}^{s}\,[c'^{s}_{ns} + c_{n}^{s} : n \in \mathrm{I}_{s}] = \mathbf{q}^{s}\,\mathbf{H}^{s}\,\mathbf{B}^{s}\,(\mathbf{c}'^{s}_{\mathrm{I}(s)} + \mathbf{c}_{\mathrm{I}(s)}^{s}). \tag{7}$$

Starting from station n, the expected cost of search and park is

$$\hat{g}_{n}^{s} = \widetilde{c}_{n}^{s} + \hat{c}_{s}(\delta_{n}^{s}, \mathbf{B}^{s}, \mathbf{P}^{s}, \mathbf{c}'^{s}_{\mathrm{I}(s)}, \mathbf{c}_{\mathrm{I}(s)}^{s}) = \delta_{n}^{s}\,\mathbf{H}^{s}\,(\overline{\mathbf{B}}^{s}\vec{\mathbf{P}}^{s}\,\mathbf{c}_{\mathrm{T}}^{s} + \mathbf{B}^{s}\,(\mathbf{c}'^{s}_{\mathrm{I}(s)} + \mathbf{c}_{\mathrm{I}(s)}^{s})). \tag{8}$$

Demand functions. On the mode layer, the set of routes R_{ℓ} of a given leg $\ell \approx (n,s)$ includes any elementary path r from initial station n to any $i \in \mathrm{I}_{s}$: here, a path is a pair sequence – leg, in which the leg part enables one to further specify a path instance. Denote the cost of path r by g_{r} and its extension to include the search and park cost up to s by $\hat{g}_{r}^{\ell} \equiv g_{r} + \hat{g}_{i}^{s}$. The assignment of customers to target stations of minimum cost is stated as follows: Find vector $\mathbf{f}_{\mathrm{LR}} = [f_{r}^{\ell} : r \in \mathrm{R}_{\ell}, \ell \in \mathrm{L}]$ and dual variables $[\psi_{\ell} : \ell \in \mathrm{L}]$ such that

$$f_{r}^{\ell} \geq 0 \quad \forall r \in \mathrm{R}_{\ell},\ \forall \ell \in \mathrm{L}, \tag{9a}$$

$$\textstyle\sum_{r \in \mathrm{R}_{\ell}} f_{r}^{\ell} = x_{\ell} \quad \forall \ell \in \mathrm{L}, \tag{9b}$$

$$\hat{g}_{r}^{\ell} - \psi_{\ell} \geq 0 \ \text{ and } \ f_{r}^{\ell}(\hat{g}_{r}^{\ell} - \psi_{\ell}) = 0 \quad \forall r \in \mathrm{R}_{\ell},\ \forall \ell \in \mathrm{L}. \tag{9c,d}$$

At solution, ψ_{ℓ} is equal to the minimum leg cost among the route options of the leg.

Station demand $\mathbf{q}^{s}$ stems from route flows in a straightforward way:

$$q_{i}^{s} = \textstyle\sum_{n \in \mathrm{I}, r \in \mathrm{R}(n,i)} f_{r}^{\ell \approx (n,s)} \ \text{ for } i \in \mathrm{I}_{s}. \tag{10}$$

The interplay of model layers. Conditions (9) and (10) relate the mode layer to the multimodal network layer: the latter supplies the former with entry-exit flows $\mathbf{x}_{\mathrm{L}}$, whereas the former supplies the latter with leg costs ψ_{L}.

The station and the mode layers interplay on the egress side of the VSS: the station layer

provides egress wait times w_n^+ and dock shortage probabilities α_n^+, while the mode layer induces the egress flows, μ_n, and the rates of acceptance to wait for egress, s_n^s, as follows:

$$\mu_n = \sum_{s\in \mathrm{I}, \mathrm{I}_s \ni n} y_n^s , \tag{11}$$

$$s_n^s = \Pr\{\text{Wait at } n \text{ is better than diversion}\} . \tag{12}$$

A Discrete Choice Model can be associated to s_n^s, e.g. a multinomial logit model with options in I_s, option disutility either $w_n^+ + c'_{ns}$ or $\tau_{mn} + \hat{g}_n^s$ for $m \in \mathrm{I}_s \setminus n$ and parameter θ, yielding

$$s_n^s = \exp[-\theta(w_n^+ + c'_{ns})] / [\exp(-\theta(w_n^+ + c'_{ns})) + \sum_{m\in \mathrm{I}_s \setminus n} \exp(-\theta(\tau_{mn} + \hat{g}_n^s)) . \tag{12a}$$

$$s_n = (\sum_{s\in \mathrm{I}, \mathrm{I}_s \ni n} y_n^s s_n^s) / \mu_n . \tag{12b}$$

4 Multimodal network

Multimodal integration. As the VSS is a mode of public transport, it can be integrated to the multimodal network as a special kind of transit service on the basis of the modal legs between the stations of access and exit. Then, the optimal strategy treatment of a leg-based transit network [DCF93], extended to several features of congestion about passengers and vehicles by [LCP12], can be adapted to our multimodal network subject to one major change: put simply, the frequency of a transit service is replaced by the Probability of Immediate Availability of a vehicle at an entry station. This requires to specify further the notion of a local travel strategy: a general theory of modal availability for traffic assignment to a multimodal network has been developed in a companion paper [Leu13], of which a simple application is presented hereafter.

Modal availability and travel strategy. At the boarding node of a VSS station n, the VSS mode provides a travel option to get to the destination z with a given quality of service in terms of minimum run cost $g_{nz}^{\mathrm{I}} = \min_{s\in \mathrm{I}} \psi_{ns} + g_{sz}$, of wait time w_n^- if no vehicle is available immediately and of the probability of unavailability, α_n^-. Alternative options have an initial link by Walk. A travel strategy involves either VSS or Walk or Combined, depending on the VSS minimum cost, g_{nz}^{I}, and the Walk average cost, g_{nz}^{W}. If $g_{nz}^{\mathrm{I}} > g_{nz}^{\mathrm{W}}$ then the Walk option is selected on a full basis; otherwise the VSS option is chosen at least with its probability of immediate availability, $1-\alpha_n^-$, and eventually on a full basis if $w_n^- + g_{nz}^{\mathrm{I}} < g_{nz}^{\mathrm{W}}$. The pure Walk strategy is a "continuous strategy" in the theory of modal availability [Leu13], whereas the pure VSS is a singleton "discrete sequence" and the "VSS or Walk" is a "hybrid strategy".

For the two problems of Finding an optimal local strategy and Finding a hyperpath heading to a given destination, there exists an optimal solution and efficient algorithms have been provided [Leu13]. On the upper layer network made of the VSS legs and the pedestrian

links, the outcome for each destination z consists in a vector of link proportions $h_a^z \in [0,1]$ such that $\sum_{a\in \mathrm{A}_n^+} h_a^z = 1$. Inefficient walk links or VSS legs have $h_a^z = 0$, while the links belonging to an optimal strategy can get a positive share. The walk modal share is split between the walk links of identical minimum cost. Similarly, the VSS modal share is split between the VSS legs of minimum run cost.

Loading the hyperpath with trip flows from origin nodes $o \in \mathrm{N}$ to a given destination node $z \in \mathrm{Z}$ is performed in the classical way on the basis of the link proportions h_a^z and in the order of decreasing cost g_{nz}^* to the destination.

The interplay of model layers. Figure 1 depicts a multimodal network: the intermediary level of pedestrian nodes and links is essential for access to and from the VSS mode, via the land nodes associated to the VSS stations. However Figure 2 depicts the model layers better: each public mode either VSS or a transit line is detailed at the lower level, whereas the upper layer is comprised only of pedestrian and modal legs.

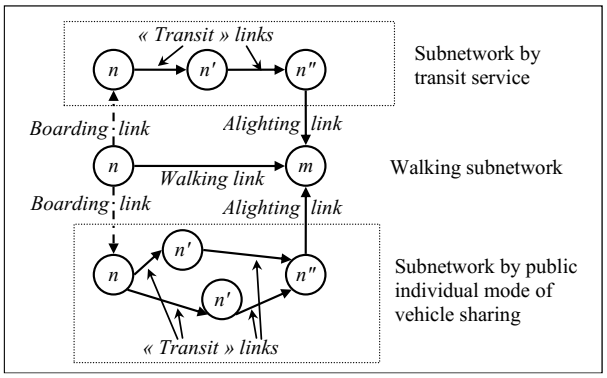


Figure 1: Multimodal network: topology of nodes, links and modes - taken from [Leu13].

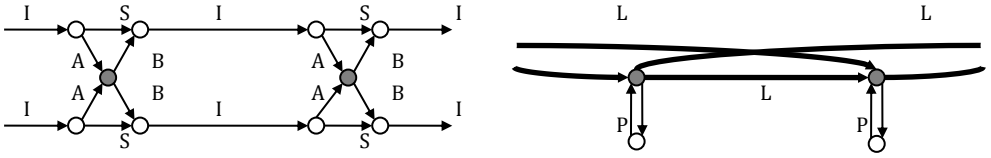


Figure 2: Bi-layer organization: (a) lower level, (b) upper layer- taken from [LCP12].

On the upper layer, the flow state is a vector of flows by destination and link, $\mathbf{x}_{\mathrm{AZ}} = [x_a^z : a \in \mathrm{A}, z \in \mathrm{Z}]$ with Z the set of destinations and A that of upper layer links. Based on $\mathbf{x}_{\mathrm{AZ}}$, the passenger flow arriving at land node n' and considering to enter the VSS mode at n towards destination z is $x_{nz}^+ = Q_{nz} + \sum_{a\in \mathrm{A}_n^-} x_{az}$ with Q_{nz} the origin-destination flow from n to z. Let $\omega_{nz} \equiv 1$ if VSS is attractive at n i.e. if $g_{nz}^{\mathrm{I}} \le g_{nz}^{\mathrm{W}}$, or 0 otherwise. In the latter case the VSS gets no flow for z and $\lambda_{nz} = 0$. If VSS is attractive, let $\tilde{\alpha}_{nz} = \overline{\alpha}_n^-$ if $g_{nz}^{\mathrm{I}} < g_{nz}^{\mathrm{W}}$ or $\tilde{\alpha}_{nz} \in [0, \overline{\alpha}_n^-]$ if $g_{nz}^{\mathrm{I}} = g_{nz}^{\mathrm{W}}$. If $w_n^- + g_{nz}^{\mathrm{I}} < g_{nz}^{\mathrm{W}}$ let $h_{nz}^{\mathrm{I}} = 1$ or $\in [\overline{\alpha}_n^-, 1]$ in case

of equality. The proportion of customers assigned to VSS is $h^{\mathrm{I}}_{nz}\omega_{nz}$, within which $\widetilde{\alpha}_{nz}\omega_{nz}$ can board immediately and the remaining part has to wait, hence $\mathrm{r}^z_n \equiv (h^{\mathrm{I}}_{nz} - \widetilde{\alpha}_{nz})/h^{\mathrm{I}}_{nz}$. It must hold that:

$$\lambda_n = \sum_{z\in Z}\lambda_{nz}\,. \tag{13}$$

$$\mathrm{r}_n = (\sum_{z\in Z}\lambda_{nz}\,\mathrm{r}^z_n)/\lambda_n\,. \tag{14}$$

5 Traffic equilibrium

Model integration. The three sub-models constituted so far, namely VSS Station, VSS Mode and Multimodal Network, make building blocks in an integrated model. Each of the building blocks is shown in an input-output setting in figure 3: their integration would give rise to a block diagram. A fourth model can be identified at that stage: that of roadway traffic. This traffic involves the two kinds of vehicle flow that pertain to the VSS mode, respectively initial flow from entry to target station ($\mathbf{f}_{\mathrm{LR}}$) and cruising flow in transition between egress stations ($\mathbf{x}^s_{\mathrm{T}}$ for each exit station s); it also involves the plain roadway traffic of private cars and duty vehicles. On each roadway link a, the local flow v_a determines the individual travel time t_a on the basis of a link travel time function as follows:

$$t_a = \mathrm{T}_a(v_a)\,. \tag{15}$$

It would be easy to model further the travel mode by private car or by any other private vehicle and to integrate it within our multimodal framework as an additional model block. For simplicity, let us use condition (15) only to represent the model of roadway traffic.

State vector. In the integrated model, the endogenous variables can be synthesized on the basis of a four-fold state vector made of $\mathbf{x}_{\mathrm{AZ}}$, $\xi^{\mathrm{N}}_{\Sigma\mathrm{Z}}$, α^{-}_{I} and η_{LR}, in which:

- $\xi^{\mathrm{N}}_{\Sigma\mathrm{Z}} = [\xi^n_{\sigma z} : n\in\mathrm{N}, z\in\mathrm{Z}, \sigma\in\Sigma(n)]$ with $\xi^n_{\sigma z}\geq 0$ and $\sum_{\sigma\in\Sigma(n)}\xi^n_{\sigma z} = 1$ is the vector of strategy proportions on the upper layer network, as in [Leu13],
- $\eta_{\mathrm{LR}} = [\eta_{\ell r} : \ell\in\mathrm{L}, r\in\mathrm{R}_\ell]$ with $\eta_{\ell r}\geq 0$ and $\sum_{r\in\mathrm{R}_\ell}\eta_{\ell r} = 1$ is the vector of flow proportions to elementary paths serving leg $\ell\approx(n,s)$ on the VSS mode (with path head in I_s),

The ξ and η vectors are required to share the demand flows among alternative strategies or alternative modal paths, respectively.

Definition of traffic equilibrium. A state vector $\mathbf{X}\equiv(\mathbf{x}_{\mathrm{AZ}}, \xi^{\mathrm{N}}_{\Sigma\mathrm{Z}}, \eta_{\mathrm{LR}}, \alpha^{-}_{\mathrm{I}})$ is a traffic equilibrium if and only if it solves the system of conditions (1)-(15).

More rigorously, a condition about strategy proportions should be stated on the basis of

condition (9) about paths, by replacing f_r^ℓ by $\eta_{\ell r}$ and by introducing a function of strategy cost.

VIP characterization. In this setting, the state vector belongs to an admissible set: a finite dimensional, bounded polytope which is nonempty and compact. The derivation of all remaining model variables from the state vector and the system of conditions is straightforward. It is also straightforward to associate a specific cost function to every strategy proportion: this cost function and the path cost function are continuous with respect to $\mathbf{X}$. It is somewhat less straightforward but still easy to associate continuous cost functions to the $\mathbf{x}_{\mathrm{AZ}}$ and α_{I}^- components (cf. [Leu13] as regards the $\mathbf{x}_{\mathrm{AZ}}$), so as to restate the traffic equilibrium with respect to the state vector and in such a way that the cost function is continuous. From this stems a characterization theorem of traffic equilibrium as a Variational Inequality problem. Then, for a solution to exist, it is sufficient to check that the conditions of local compatibility are feasible within the admissible set.

Existence of equilibrium and the issue of feasibility. The compatibility conditions pertain to the domain of vehicle shortage, $r < \mu/\lambda$, and that of dock shortage, $s < \lambda/\mu$. When solving for equilibrium, based on the current value of $\mathbf{X}$, the r_n rate at n stems from passengers willing to wait for a vehicle rather than to walk to a subsequent node, because the wait time is not so high. As $w_n^- = 1/(\mu_n - r_n\lambda_n)$, it increases with r_n. If $\mu_n > \lambda_n$ then the compatibility condition is satisfied, otherwise w_n^- can take values as large as required by increasing r_n, thus reducing the attractivity of the wait option and leading to reduce r_i: so a fixed point must exist which ensures the compatibility.

On the egress side, a similar argument applies: if $\lambda_n > \mu_n$ then the compatibility requirement about dock shortage is satisfied. Otherwise, as $w_n^+ = 1/(\lambda_n - s_n\mu_n)$ is an increasing function of s_n, by increasing w_n^+ it comes out that the wait option will be less demanded than the diversion to neighboring stations. In turn, this induces a lower s_n, so a fixed point must exist, which ensures the compatibility.

Equilibration algorithm. As the VIP is endowed with a regular cost function, it can be solved by successive approximations: at each iteration in this process, a related yet simpler problem called auxiliary is solved to yield an auxiliary state which is used to enhance the current state. This Auxiliary Program Principle as developed by Guy Cohen (1984) is well-known in assignment theory in the two instances of the Frank-Wolfe algorithm and the Method of Successive Averages (MSA). A typical step would proceed as follows. At iteration i, given current state $\mathbf{X} = (\mathbf{x}, \xi, \alpha, \eta)$, evaluate $\mathbf{F}(\mathbf{X})$ and solve the VIP with cost function $\mathbf{F}(\mathbf{X})$, which amounts to minimize the duality gap function $\mathbf{F}(\mathbf{X}).(\mathbf{X} - \widetilde{\mathbf{X}})$. This yields an auxiliary state $\widetilde{\mathbf{X}} = (\widetilde{\mathbf{x}}, \widetilde{\xi}, \widetilde{\alpha}, \widetilde{\eta})$. From $\mathbf{X}$ and $\widetilde{\mathbf{X}}$, the next current state $\mathbf{X}'$ is derived primarily by relaxation i.e. $\mathbf{X}' \equiv \mathbf{X} + \zeta_i(\widetilde{\mathbf{X}} - \mathbf{X})$ wherein $(\zeta_i)_{i\geq 0}$ is a decreasing sequence of positive numbers converging to zero sufficiently quickly. Some caution must be exerted to maintain

an acyclic lattice of optimal strategies on the upper layer network, cf. [LCP13].

However it would be computationally awkward to deal with the ξ and η parts explicitly. An alternative, simpler approach is to focus on $\mathbf{x}_{\mathrm{AZ}}$ on the upper layer, on α_{I}^{-} for the VSS stations and on a vector of link flows on the VSS lower layer, say $\mathbf{v}_{\mathrm{A(I)}}$, and to handle them jointly by a three-fold Method of Successive Averages.

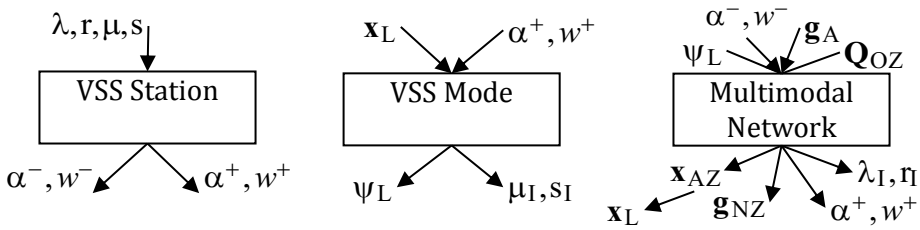


Figure 3: Sub-models as input-output functions.

6 Toy instance

A VSS mode with three stations $\mathrm{I}=\{0,1,2\}$ includes (0) a central station, (1) a first ring lot, (2) a second ring lot. Run cost by VSS, excluding exit cost, amounts to γ between 2 and 1 or 1 and 0 and to 2γ between 2 and 0, whereas by walk it amounts to $\widetilde{\gamma}$ or $2\widetilde{\gamma}$. Figure 4 depicts the lower-layer and upper-layer networks: assuming a single O-D pair from 2 to 0 and two exit nodes $s\in\{0,1\}$ with same catchment area $\mathrm{I}_s=\{0,1\}$, only three legs can be used on the upper layer: $\mathrm{L}=\{(2,0),(2,1),(1,0)\}$.

Here our objective is restricted to illustrate the formation of travel costs along the VSS node. By exit station s and egress node $n\in\mathrm{I}_s$ as "initial target", the stopping probability is β_n^s and the remaining users are diverted to node $1-n$, yielding that

$$\overline{\mathbf{B}}^s=\begin{bmatrix}\overline{\beta}_0^s & 0\\ 0 & \overline{\beta}_1^s\end{bmatrix},\ \mathbf{P}^s=\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}\ \text{and}\ \mathbf{H}^s=(\mathbf{U}^s-\overline{\mathbf{B}}^s\mathbf{P}^s)^{-1}=\frac{1}{\Delta_s}\begin{bmatrix}1 & \overline{\beta}_0^s\\ \overline{\beta}_1^s & 1\end{bmatrix}\ \text{with}$$

$\Delta_s=1-\overline{\beta}_0^s\overline{\beta}_1^s$.

The vector of transition costs is $\mathbf{c}_{\mathrm{T}}^s=[0\quad\gamma\quad\gamma\quad 0]^t$. The exit cost at n for exit station s is $\hat{g}_n^s=\widetilde{c}_n^s+\hat{c}_n^s$, wherein the search cost from n is $\widetilde{c}_n^s=\gamma\overline{\beta}_n^s(1+\overline{\beta}_{1-n}^s)/\Delta_s$ and the final cost at n is $\hat{c}_n^s=[(\beta_n^s-\overline{\alpha}_n^+)w_n^+ + \overline{\beta}_n^s(\beta_{1-n}^s-\overline{\alpha}_{1-n}^+)w_{1-n}^+ + \widetilde{\gamma}\overline{\beta}_{1-n}^s(\delta_{sn}\beta_n^s+\delta_{s,1-n})]/\Delta_s$. By modal path r serving leg $\ell\approx(i,s)$, the modal cost including run cost and exit cost is as follows:

- $\ell\approx(1,0)$: path 1-0 costs $g_{1-0}^{(1,0)}=\gamma+\hat{g}_0^0$.
- $\ell\approx(2,0)$: path 2-0 costs $g_{2-0}^{(2,0)}=2\gamma+\hat{g}_0^0$ and path 2-1 costs $g_{2-1}^{(2,0)}=\gamma+\hat{g}_1^0$.
- $\ell\approx(2,1)$: path 2-0 costs $g_{2-0}^{(2,1)}=2\gamma+\hat{g}_0^1$ and path 2-1 costs $g_{2-1}^{(2,1)}=\gamma+\hat{g}_1^1$.

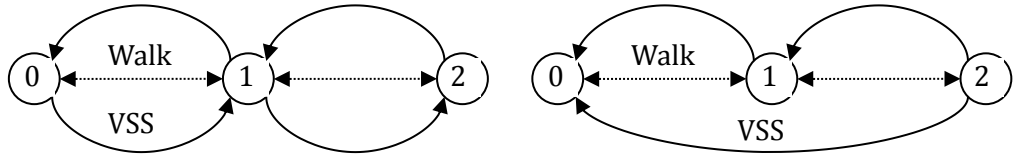


Figure 4: Instance network: (a) lower level, (b) upper layer.

7 Conclusion

A framework has been provided to model travel demand for a Vehicle Sharing System within a multimodal setting, in both competition and complementariness to alternative modes. The basic principle is to treat the VSS mode as some kind of public service which provides station-to-station legs to the trip-makers. The service egress involves to park the vehicle at a capacitated station, eventually with cruising between stations. The service access involves the issue of availability at an entry station. The integrated model is comprised of building blocks by station and travel mode. The trip-makers' paths are embedded into hyperpaths on the upper layer network but may involve loops for parking on the lower layer of VSS mode.

By its modular construct, the model can be easily extended to include transit services as well as individual modes using private vehicles. Work is going on at LVMT to implement a model software and apply it to the Paris metropolitan area. After consolidating the simulation capability, it is planned to test a range of transportation policies, including capacity development, fleet sizing and tariff setting for the Vehicle Sharing System.

References

[Cia13] F. CIARIA, N. SCHUESSLER, and K. W. AXHAUSEN: "Estimation of Carsharing Demand Using an Activity-Based Microsimulation Approach: Model Discussion and Some Results". In: *International Journal of Sustainable Transportation* 7.1 (Jan. 2013), pp. 70–84. URL: http://www.tandfonline.com/doi/pdf/10.1080/15568318.2012.660113.

[LCP12] F. LEURENT, E. CHANDAKAS, and A. POULHÈS: "A Passenger Traffic Assignment Model with Capacity Constraints for Transit Networks". In: *Elsevier Procedia - Social and Behavioral Sciences* 54 (Oct. 2012), pp. 772–784.

[LeB12] F. LEURENT and H. BOUJNA: "Traffic Equilibrium in a Network Model of Parking and Route Choice, with Search Circuits and Cruising Flows". In *Elsevier Procedia - Social and Behavioral Sciences* 54 (Oct. 2012), pp. 808–821.

[Leu12] F. LEURENT: "Modelling a vehicle-sharing station as a dual waiting system: stochastic framework and stationary analysis". In: *EURO Transportation and Logistics Journal*. submitted on Nov. 22, 2012. URL: http://hal.archives-ouvertes.fr/hal-00757228.

[Leu13] F. LEURENT: *On modal availability, travel strategies and traffic equilibrium on a multimodal network*. Enpc working paper. URL: http://hal.archives-ouvertes.fr/hal-00827631

Corresponding author: Fabien Leurent, Université Paris-Est, Laboratory on City, Mobility and Transportation, Ecole des Ponts ParisTech – IFSTTAR – UPEM, 77455 Champs sur Marne, France, phone: +00 33 181 668 854, e-mail: fabien.leurent@enpc.fr

Table of Authors

Abreu, Giuseppe, 243
Albrecht, Thomas, 343, 480
de Almeida, David , 449
Asamer, Johannes, 34
Axelsson, Stefan, 154

Baiocchi, Andrea, 111
Bäker, Bernard, 68
Bešinović, Nikola, 459
Bifulco, Gennaro N., 233, 513
Binder, Anne, 480
Blieberger, Johann, 354
Bolic, Tatjana, 123
Bommes, Michael, 223
Born, Alexander, 165
Brands, Ties, 301
Brockfeld, Elmar, 271
Buisson, Christine, 449

Cantarella, Giulio E., 513
Carlsson, Bengt, 154
Castaldi, Claudia, 291
Castelli, Lorenzo, 123
Cattrysse, Dirk, 387
Ciccarelli, Gennaro, 291
Colombaroni, Chiara, 111, 291
Corman, Francesco, 417, 439
Cuniasse, Pierre-Antoine, 449
Cuomo, Francesca, 111

D'Ariano, Andrea, 133, 439
D'Ariano, Paolo, 133
Dewilde, Thijs, 387
Döge, Klaus-Peter, 190
Dunkel, Juliane, 377

Esztergár-Kiss, Domokos, 46

Fazekas, Adrian, 223
Fehrenbach, Lena, 34
De Felice, Mario , 111
Friedrich, Markus, 34
Friso, Klaas, 78
Fusco, Gaetano, 111, 291

Gäbel, Charlotte, 332
Galante, Francesco, 233
Gassel, Christian, 311
Gentile, Guido, 502, 523
Gerdts, Matthias, A.144
Ghods, Alireza, 243
Gosda, Uwe, 200
Goverde, Rob M. P., 301, 417, 459, 469
Grimm, Jan, 181
GroßmannPeter, 398, 407
Guler, S. Ilgin, 321

Heilmann, Bernhard, 34
Heimgartner, Christian, 11
Herrera-Pinzón, Iván, 165
Hoogendoorn, Serge, 281
Huang, Wei, 493
Huerlimann, Daniel, 377

Jaekel, Birgit, 343

Karon, Grzegorz, 367
Kecman, Pavle, 469

Klipphahn, Samuel, 253
Klunder, Gerdien, 281
Kohlen, Ralf, 265
Kostic, Bojan, 502
Krimmling, Jürgen, 68, 311
Krumnow, Mario, 68
Kucharski, Rafał, 523
Kuhns, Günter, 271
Kümmling, Michael, 407

Labinsky, Alexander, 398
Lämmer, Stefan, 23, 56, 89
Laumanns, Marco, 377
Leurent, Fabien, 533
Lohmiller, Jochen, 34
Looser, Jonas, 377
Löwe, Stefan, 332

Maciejewski, Michał, 1
Marlière, Grégory, 428
Mein, Edwin, 78
Menendez, Monica, 11, 321
Meschini, Lorenzo, 523
Michler, Oliver, 200, 253
Mikulski, Jerzy, 367

Nachtigall, Karl, 407
Nagel, Kai, 1
Neumann, Thorsten, 271

Oeser, Markus, 223
van Oort, Niels, 301
Opitz, Jens, 398, 407
Osekowska, Ewa, 154

Pacciarelli, Dario, 133, 439
Palagachev, Konstantin D., A.144
Pariota, Luigi, 233
Pellegrini, Paola, 428
Pillat, Juliane, 34
Preusker, Michael, 332

Quaglietta, Egidio, 417, 459

Rausch, Markus, 23, 89
Reinthaler, Martin, 34
Rieck, Matthias, A.144
Rigonat, Desirée, 123
Ringel, Julia, 253
Rodriguez, Joaquín, 428, 449
Roos, Samuel, 377
Russo, Spena Maria, 233

Sabene, Federico, 439
Samà, Marcella, 133
Samà, Marcella, 439
Schedler, Karl, 34
Schöbel, Andreas, 354
Schubert, Torsten, 68
Sels, Peter, 387
Severi, Stefano, 243
Seybold, Bernhard, 377
Simonelli, Fulvio, 513
Smith, Mike J., 493
Sohr, Alexander, 271
Sparing, Daniel, 301
Steiner, Albert, 377

Taale, Henk, 281
Teboul, Emmanuel, 449
Tiddi, Daniele, 502
Tischler, Kathleen, 56
Touko, Louis, 271
Traxler, Johannes, 211
Treiber, Martin, 23, 89, 99

Vansteenwegen, Pieter, 387
Válóczi, Dénes, 46
Velonà, Pietro, 513
Viti, Francesco, 493
Volcic, Mark, 354

Weber, Richard, 200
Weiß, Reyk, 398, 407
Wuest, Raimond, 377

Zantema, Kobus, 78